

Finding gaps in the Kamasau language corpus guide

By Joy Sanders
Dallas International University



Introduction:

- Woodbury (2014): how to access language collections?
- Aim of this presentation: how can diverse audiences access the Kamasau language corpus?
- Primary focus – for the Kamasau people to access videos on YouTube and archiving them for future generations
- Secondary focus – to make quality recordings available to the academic community
- How gaps were filled in the Kamasau language corpus.

Woodbury, Anthony. 2014. “Archives & Audiences: Toward Making Endangered Language Documentation People Can Read, Use, Understand, & Admire.” *Language Documentation & Description* (London: SOAS) 12: 19–36. <https://doi.org/https://doi.org/10.25894/ldd161>.



Linguistic Background

The Kamasau language has been classified as a part of the Marienberg Hills branch of the Torricelli language family (Foley 2017, 296–304).



Four dialects (Sanders and Sanders):

Segi:
Tring and Wau

Soigi:
Kamasau

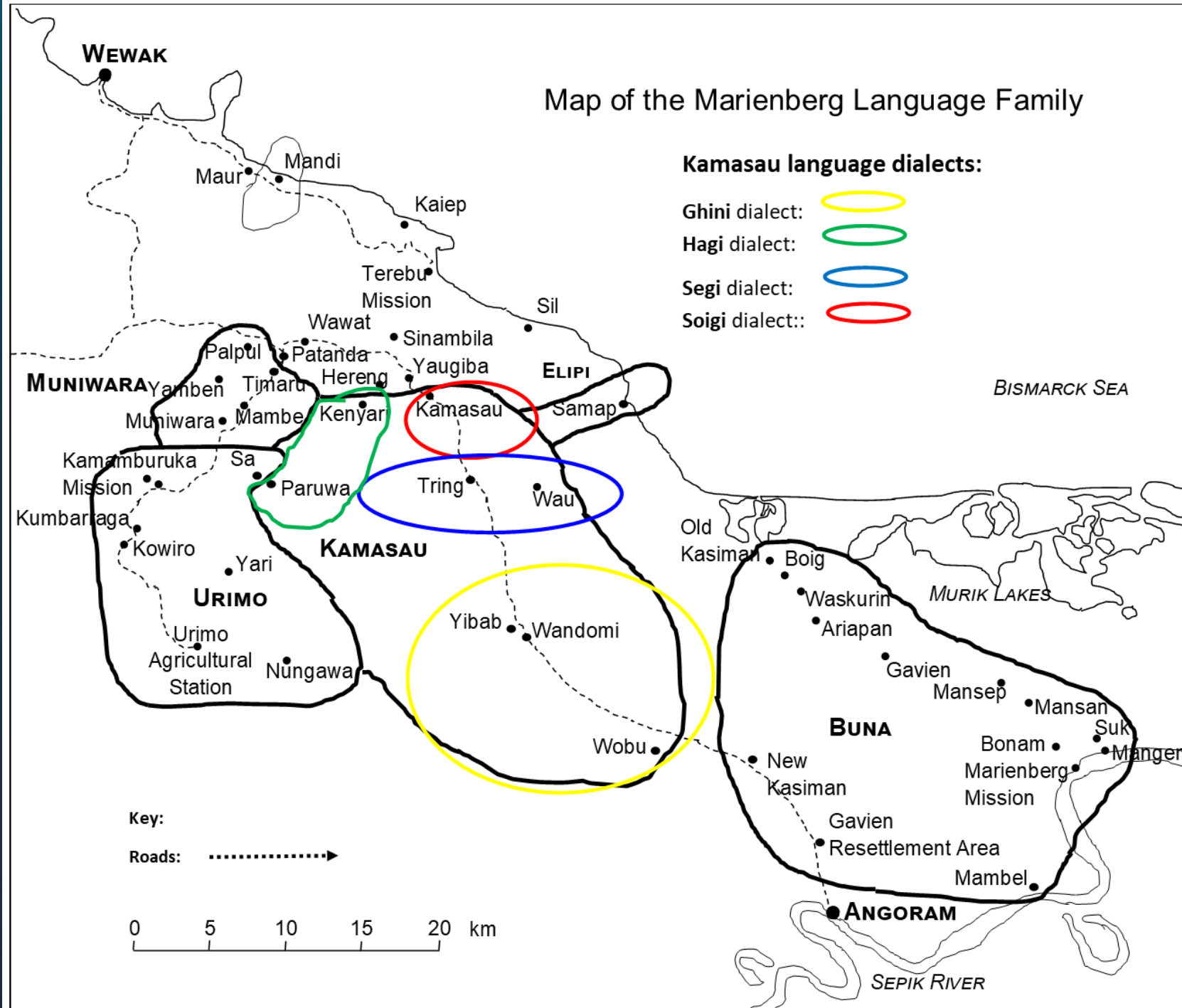
Ghini:
Ibab and Wandomi

Hagi:
Kinyare

Foley, William A. 2017. “The Languages of the Sepik-Ramu Basin and Environs.” In *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*, Vol. 4:197–432. Berlin: Walter de Gruyter.

Sanders, Joy, and Arden G. Sanders. 1980b. “Dialect Survey of the Kamasau Language.” In *Papers in New Guinea Linguistics* No. 20, 137–69. Series A, No. 56. Canberra: Pacific Linguistics.

Map of the Marienberg Language Family



1977-1999

Author & her husband worked in Tring, ESP, PNG under SIL.
Focus: language learning, linguistic analysis, literacy, and translation with several linguistic outputs.

Phonology

Orthography

Morpheme to
sentence
grammar

Anthropology

Published
Triglot
Dictionary



2002-2013

Author archived legacy data in
SIL Language and Culture Archive

Linguistic
outputs

Primers,
Readers

Books from
Writer's
Workshops

Traditional
stories

Shell books

Song Books



2013

KMS Dictionary
into FLEx

And onto
Webonary.org

- Lack of recordings noted-
- Only 65 songs
- One recorded text.
- Wrote village to ask if they wanted author to come record stories and events.



Community ownership



In 2013, the community invited the author to do video and audio recordings of the language and culture.



They planned which cultural events were to be video recorded.



Individuals prepared for recordings.



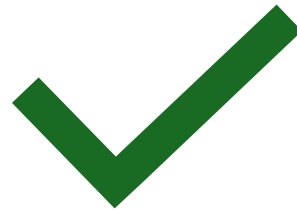
Others assisted in transcribing and translating those.



Linguistic Goals



Word lists and verb
paradigms



Confirm proposed
EGIDS level- with a
survey



Identify noun
classes in printed
dictionary



Recorded data showed allophonic variation

Voiced bilabial and velar approximants that were originally analyzed as fricatives—were consistently approximants.

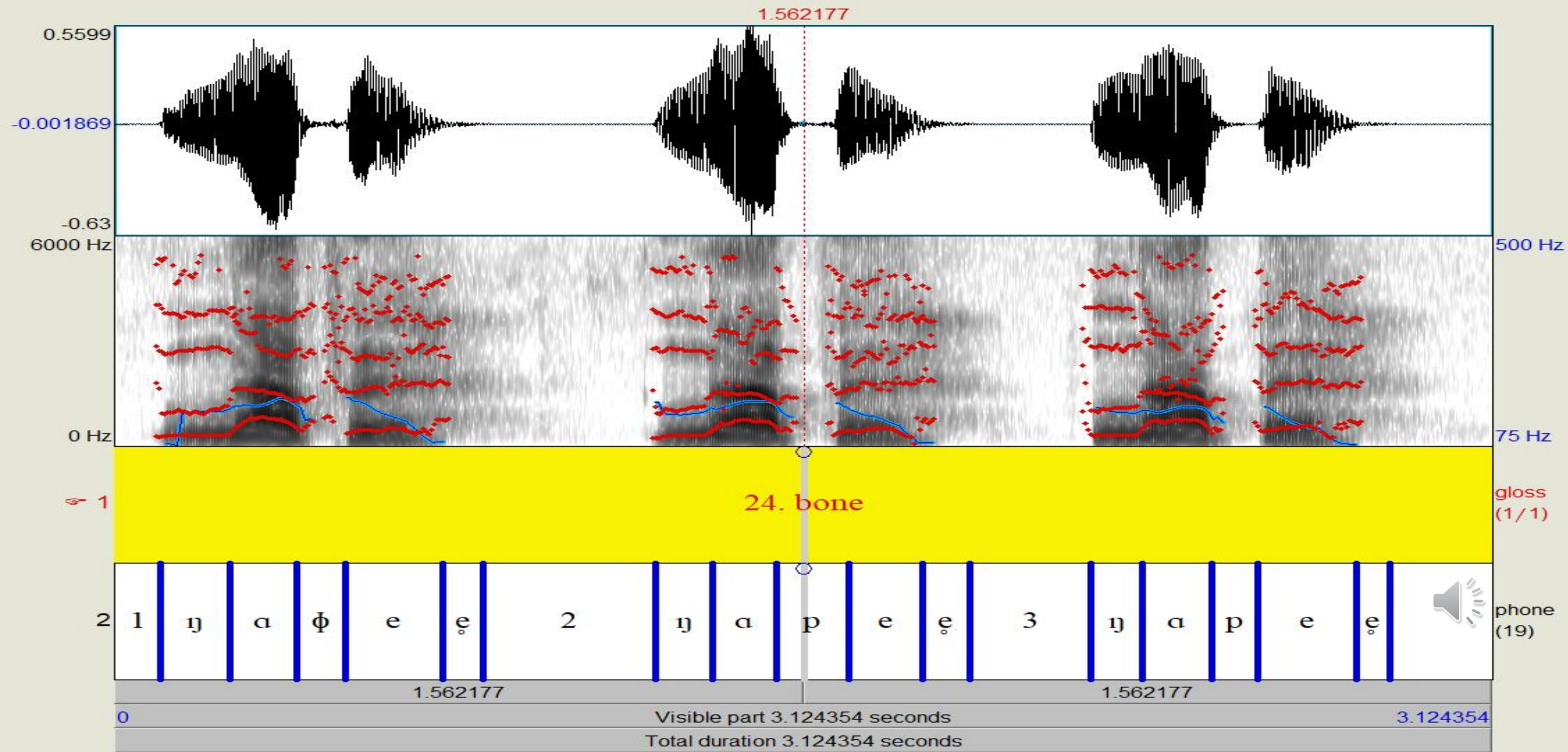
Voiceless bilabials showed variation as approximants, fricatives, affricates and stops. Acoustics confirmed the data.

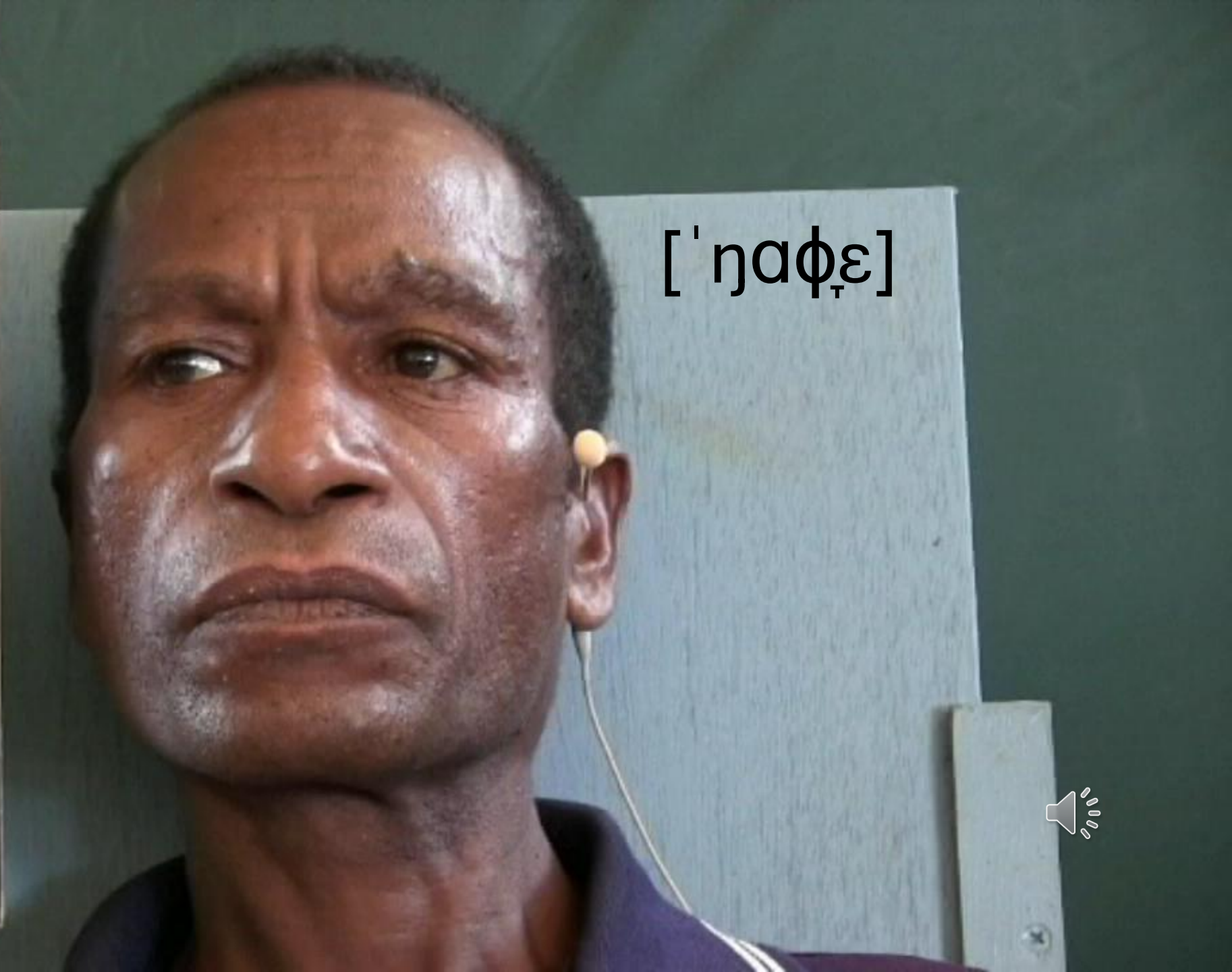


ngape
['ɲape / 'ɲaɸe]



S2: T1 ['ηαφε]; T2 & 3 ['ηαπε]

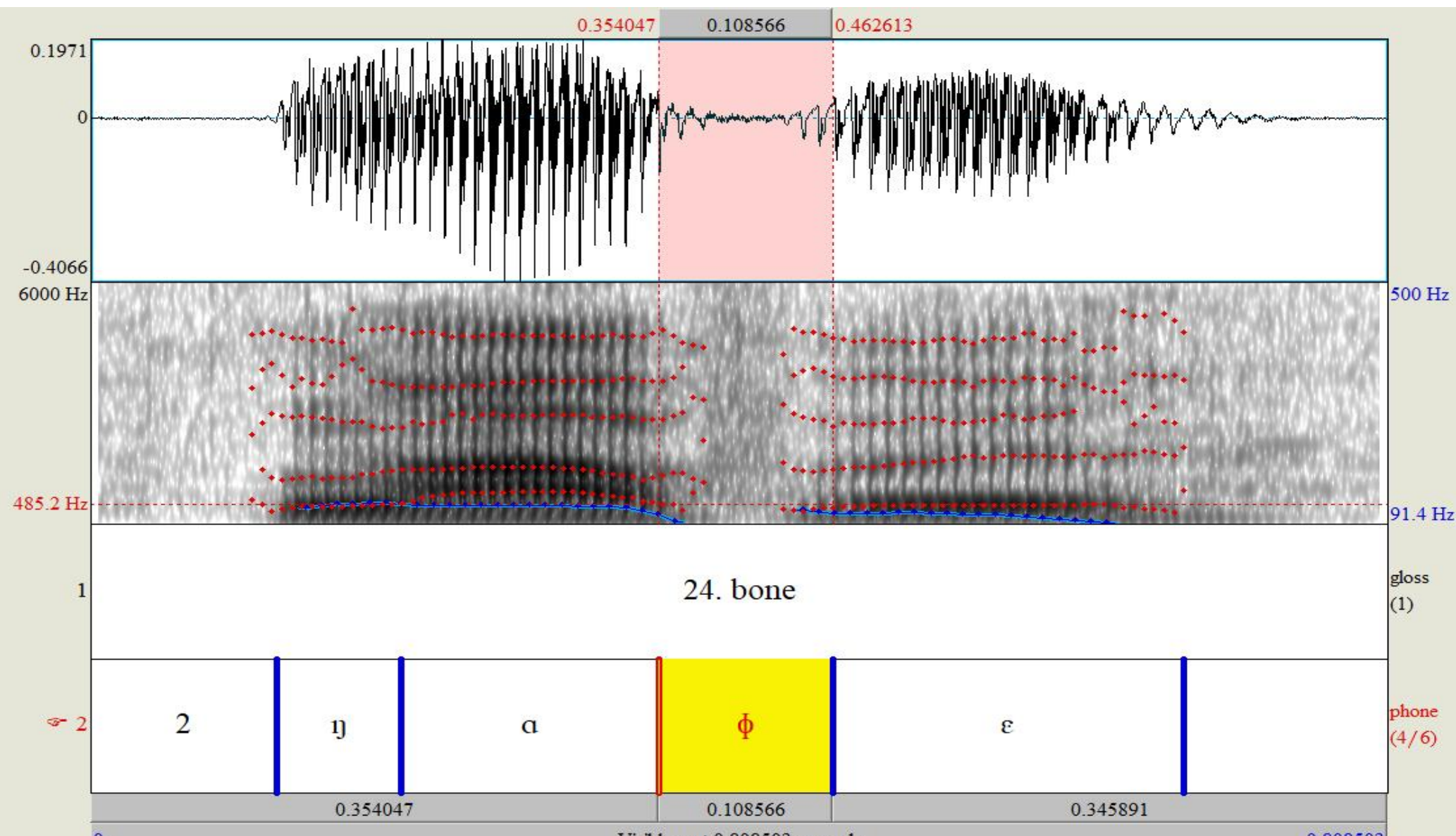




[' η α φ ε]



S1: voiceless bilabial approximant ['ɲɑɸe]



These are shown as a sample of archived data.

- 2 speakers shown were from same village, but different ages.
- Intra and inter-speaker variation was confirmed with video and acoustic data.
- Acoustic data from DIAL Conference (2022)
- These are 2 of the 5 speakers with each word list containing 191 words and 56 clauses.
- Available in SIL archive and also on YouTube:
- <https://www.youtube.com/@wandtuan6606>

Holt, Charles and Sanders, Joy. 2022. “Allophonic Variation in Voiceless Bilabials in Kamasau.” Conference Presentation. DIAL Conference, Dallas, TX, October 10. <https://diu.edu/academics/dial-conference-presentations/#allophonic-variation>.



Language & Culture videos

Sago splitting &
pounding

Sago washing &
packaging

Cooking sago in
traditional clay
pots

Making a sago
mallet

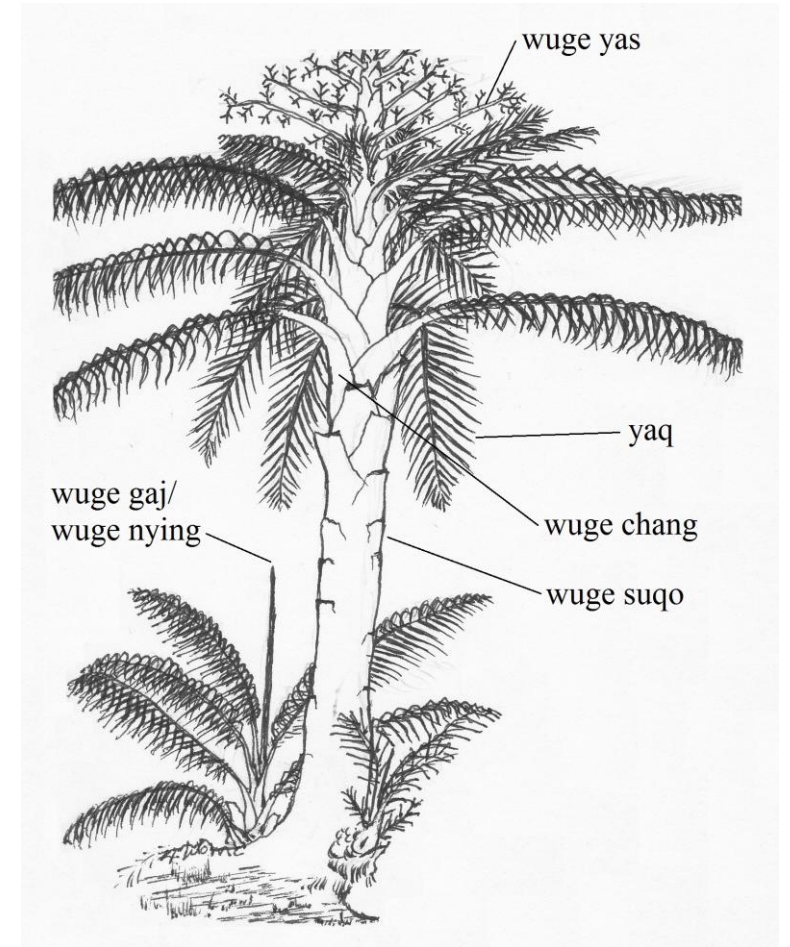
Making a slit
gong drum
(garamut)

Making a grass
skirt

Making fire
without
matches

Making a
crayfish trap

Crayfish trap
and coconut
traditional story



By John August Womosa



Sanders, Joy. 2025. "The Significance of Documenting Culture in the Kamasau [Kms] Language Documentation Project." *Language & Linguistics in Melanesia: Journal of the Linguistic Society of Papua New Guinea* 43: 32–62. https://www.langlxmlanesia.com/_files/ugd/4bb486_863889a3f87d43779789a4ab0670256a.pdf.

Making a slit gong
drum (garamut)
by Leo Amu



Aye te...
Something else...



Kamasau Language Vitality

Sociolinguistic survey in 2014 showed language is on the cusp of shifting.

Survey confirmed EGIDS* level 7, since the children are not being taught the language.

Lewis, Paul & Gary Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique* 2.

* EGIDS= Expanded Graded Intergenerational Disruption Scale.



2014

Data Management Flow

1. Field data
management
notebooks

2. Entered
in File location
spreadsheet

3. Files
organized in
SayMore

4. Files
processed in
SayMore to FLEx
and ELAN



Field data
management
notebook:
sample page

SD TR4152		Hard to hear	
140627-			
Event ID:	SD03 ← Zoom ID 101335	Date:	27/06/2014
Devices:	Canon Vixia HFG20 Sony Cyber Shot Zoom H2 Canon G10		
Mics:	Built-in Shure SM58 MMPSM-Pro mic Polaroid video mic Pro PT Mic		
Sample Rate (KHz):	44.1	Bit Depth:	16
Format:	KMS-CO1		
Channels:	Mono Stereo	Length:	
Title:	Washing sago	Informed Consent?	File 140627-100713
Speakers:	Markus Asare Ruth Gobit Dorothy Gisla		
Location:	Yangid - close to Tring, ESP, PNG		
Setting/ Situation:	Washing sago, putting sago in a form		
Photo/Vid			
Spontaneity:	Planned	Modality:	Speaking
Naturalness:	Natural setting for processing sago		
Genre:	Procedural - how to wash sago		
Synopsis:	Descriptive - description of equipment used to wash sago.		
Other Contributors:			
Name	Rec	Tscr	Tltr
Dorothy			
Joy Sanders video			
Kelsey Daniels zoom			
Sonya Paoli still			

SD TR4152		Hard to hear	
140627-			
Event ID:	110-242	Zoom ID	SD03
Date:	27/06/2014		
Devices:	Canon Vixia HFG20 Sony Cyber Shot Zoom H2 Canon G10		
Mics:	Built-in Shure SM58 MMPSM-Pro mic Polaroid video mic Pro PT Mic		
Sample Rate (KHz):	44.1	Bit Depth:	16
Format:	KMS-CO2		
Channels:	Mono Stereo	Length:	
Title:	Pounding sago	Informed Consent?	
Speakers:	Ignas Amu		
Location:	Yangid - close to Tring, ESP, PNG		
Setting/ Situation:	Sago swamp, pounding sago & splitting		
Photo/Vid			
Spontaneity:	Planned	Modality:	Speaking
Naturalness:	Natural setting for pounding sago		
Genre:	Procedural		
Synopsis:	How to split a portion of the sago trunk in order to pound it		
Other Contributors:			
Name	Rec	Tscr	Tltr
Kelsey Daniels video			
Joy Sanders zoom			
Sonya Paoli Canon G10			



F4 ▾ : ✖ ✔ fx ▾ Initiation Rights Stories									
	A	B	C	D	E	F	G	H	
1	File Location Spreadsheet								
2	Date	Card	Folder name	File name	SayMore file name	Contents	Transcribed NB	Transc in SayM	Transl.
3	07/04/14	Transcend 5094	Folder 02	Zoom0001	Ewa Marap_Consent.wav, Simon Mo	Simon and Ewa Informed Consent			
4				Zoom0002	KMS_A07_OriginalRecording.wav	Initiation Rights Stories	yes	yes	
5				Zoom0003	KMS_A07_EnglishTranslation.wav	English Translation of Initiation Rights			
6				Zoom0004	Michael Mondu_Consent.wav, KMS_	Informed Consent of Michael Mondu			
7				Zoom0005	KMS_A08_OriginalRecording.wav	Female Witch Legend	yes	yes	yes
8				Zoom0006	KMS_A08_EnglishTranslation.wav	English Translation of Female Witch Story			
9				Zoom0007	KMS_A09_OriginalRecording.wav	Male Witch Legend	yes	partially	
10				Zoom0008	KMS_A09_EnglishTranslation.wav	English Translation of the Male Witch Story			
11				Zoom0009	Matthew Mondu_Consent, KMS_A10	Matthew Mondu Informed Consent			
12				Zoom0010	KMS_A10_OriginalRecording.wav	Origin of the Kamasau People	yes	yes	yes
13				Zoom0011	KMS_A10_Translation.wav	Tok Pisin Translation of Origin Story			
14				Zoom0012	KMS_5K1_InformedConsentandWoi	Raphael Barowai Informed Consent and Word List 1			
15				Zoom0013	KMS_5K1_WordList2.wav	Raphael Barowai Word List 2			

File Location Spreadsheet



SayMore Screenshot of Sago splitting and pounding session

Segi Texts - SayMore 3.3.0 (Release)

Project Session Person Help

Project Sessions People

Sessions

Id	Title
KMS_A26	Gisla's Testimony
KMS_A27	Gisla's Evil Spirit Story
KMS_A28	Was Nait
KMS_A29	Going to the Garden
KMS_A30	Going to Sago
KMS_A31	Arden's Parents and Sago
KMS_A32	Gisla's Prayer
KMS_A33	PMV Story
KMS_C00	Welcome Singing
KMS_C01	Washing Sago
KMS_C02	Pounding Sago
KMS_C03	Fish Trap
KMS_C04	Gertrude's Grass Skirt
KMS_C05	Making Fire
KMS_C06	Comb
KMS_C07	Cooking Soup Sago
KMS_C08	Working Bilum and String
KMS_C09	String Tricks
KMS_C10	AOG Church Service
KMS_C11	Sago Mallet
KMS_C12	Jew's Harp
KMS_C13	String Tricks Take 2
KMS_C14	Kamasau Welcome and Prayer
KMS_C15	Garamut
KMS_C16	Birds
KMS_C17	Cecilia's Grass Skirt
KMS_C18	Tok Ples Church
KMS_C19	Clay Pot
KMS_C20	Death in the Village
KMS_C21	Flute
KMS_C22	Meeting about Oil Palm
KMS_C23	Tok Ples Church
KMS_C24	Bung Kai
KMS_C25	Kamasau Farewell
KMS_C26	Leo's Garamut
KMS_C27	Revival Church Service

Open ▾ Rename... Convert... Add Files...

Name	Type	Date Modified	Size
KMS_C02.session	Session	7/30/2026 11:19 ...	
KMS_C02_OriginalRecording1.mp3	Audio	4/9/2019 12:37 PM	1.22
KMS_C02_OriginalRecording1.WAV	Audio	6/27/2014 11:04 ...	13.4
KMS_C02_OriginalRecording1.WAV.annotations.eaf	Annotations	4/9/2019 2:46 PM	1
KMS_C02_OriginalRecording1.WAV.annotations.eaf.001	001 File	8/21/2015 4:35 PM	1
KMS_C02_OriginalRecording1.WAV.flextext	FLEXTEXT File	5/7/2015 1:53 PM	
KMS_C02_OriginalRecording2.mp3	Audio	4/9/2019 12:38 PM	7.12
KMS_C02_OriginalRecording2.WAV	Audio	6/27/2014 11:12 ...	78.4
KMS_C02_OriginalRecording2.WAV.annotations.eaf	Annotations	10/3/2016 9:29 AM	4
KMS_C02_OriginalRecording2.WAV.annotations.eaf.001	001 File	9/29/2016 2:50 PM	4
KMS_C02_OriginalRecording2.WAV.flextext	FLEXTEXT File	5/7/2015 1:49 PM	3
KMS_C02_OriginalVideo.MTS	Video	6/26/2014 8:12 PM	1.7
KMS_C02_Origwav1_freeTranslation_subtitle.srt	SRT File	8/21/2015 3:11 PM	
KMS_C02_Origwav1_transcription_subtitle.srt	SRT File	4/9/2019 2:50 PM	8
KMS_C02_Origwav2_freeTranslation_subtitle.srt	SRT File	4/9/2019 2:52 PM	

Session Status & Stages Contributors Notes

ID
KMS_C02

Date
6/27/2014

Title
Pounding Sago

Setting
Yangid

People
Ignas Amu (speaker); Hespon Goi (participant); Hespon Goi (speaker)

Location
Close to Tring Village

Genre
Procedural Disc

Access

Situation

Description

- Field Language Explorer (FLEX)
- sample verb morpheme analysis:
- Verb go

Text				
Title		Kam		
Eng		KMS_5V1_Go.WAV		
Info	Baseline	Gloss	Analyze	Tagging
Print View	Text Chart			
4	Word	Ni	no	.
	Morphemes	ni	n-	o
	Lex. Entries	ni	n-	o
	Lex. Gloss	3	3.m.sg	go
	Lex. Gram. Info.	pro	v:SUBJ	vi
	Word Gloss	he	he goes	
Free He (the man) goes.				
5	Word	Ni	wuso	.
	Morphemes	ni	wu-	uso
	Lex. Entries	ni	wu-	o
	Lex. Gloss	3	3.f.sg	go
	Lex. Gram. Info.	pro	v:SUBJ	vi
	Word Gloss	she	she goes	
Free She (the woman) goes.				
6	Word	Nyombui	wo	kuso
	Morphemes	nyombui	wo	ku- so
	Lex. Entries	nyombui	wo	ku ₂ o
	Lex. Gloss	dog	child	3.n.sg go
	Lex. Gram. Info.	n neut	n neut	v:SUBJ vi
	Word Gloss	dog	baby	it goes
Free It (the little puppy) goes.				
7	Word	Wand	viso	.
	Morphemes	wand	v-	iso
	Lex. Entries	wand	v-	o
	Lex. Gloss	talk, speech	3.ling.sg	go
	Lex. Gram. Info.	n neut	v:SUBJ	vi
	Word Gloss	talk	it goes	
Free It (the vernacular language) goes.				



ELAN file: carry ditransitive/benefactive (plural benefactive)

ELAN 7.1 - KMS_5V1_CarryDitransitivePL.interlinear.eaf

File Edit Annotation Tier Type Search View Options Window Help

Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls

Volume: 100

0 50 100

KMS_5V1_CarryDitransitivePL.WAV

Mute Solo

Rate: 100

0 100

00:04:46.778 Selection: 00:00:00.000 - 00:00:00.000 0

Selection Mode Loop Mode

KMS_5V1_Carry...

00:04:32.000 00:04:33.000 00:04:34.000 00:04:35.000 00:04:36.000 00:04:37.000 00:04:38.000 00:04:39.000 00:04:40.000 00:04:41.000 00:04:42.000 00:04:43.000 00:04:44.000 00:04:45.000

00:04:32.000 00:04:33.000 00:04:34.000 00:04:35.000 00:04:36.000 00:04:37.000 00:04:38.000 00:04:39.000 00:04:40.000 00:04:41.000 00:04:42.000 00:04:43.000 00:04:44.000 00:04:45.000

A_phrase-txt-km [49] Nungoqi wurauqu/wurauq! Nungoqi wuram! Nungoqi wurany!

A_phrase-gls- [49] You all carry it for yourselves! You all carry it for them, the men! You all carry it for them, the women!

A_phrase-seg [48] 46 47 48

A_word-txt-km [159] Nungoqi wurauqu / wurauq ! Nungoqi wuram ! Nungoqi wurany !

A_morph-txt- [276] nungoqi w- *ura -qu w- *ura -uq nungoqi w- *ura -m nungoqi w- *ura -ny

A_morph- [276] nungoqi w- *er -qu w- *er -qu nungoqi w- *er -m nungoqi w- *er -ny

A_morph-g [276] 2.pl 2.pl.im carry f, 2.pl.be 2.pl carry f, 2.pl.be 2.pl 2.pl.imp carry for, 3.m.pl.b 2.pl 2.pl.i carry f, 3.f.pl.b

A_morph-h [47] 1 1 1 1

A_morph- [276] pro v:SUB vd vd:BE pro v:SUBJ vd vd:BEN pro v:SUBJ vd vd:BE

A_morph-t [276] stem prefix bound suffix prefix bound suffix stem prefix bound s, suffix stem prefix bound suffix

A_word-gls- [104] you pl you all carry for you a you all carry for you a you pl you all carry for them (m) you pl you all carry for them

Literature Review

- Himmelman (1998)- focus on primary data rather than analysis
- Woodbury (2014)- organize the data by individual speech events- include adequate metadata
- Thieberger (2016)- early proponent of language documentation- get language data archived!

- Himmelmann, Nikolaus. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36: 165–76.
hrelp.org/events/workshops/eldp2005/reading/himmelman.n.pdf.
- Woodbury, Anthony. 2014. "Archives & Audiences: Toward Making Endangered Language Documentation People Can Read, Use, Understand, & Admire." *Language Documentation & Description* (London: SOAS) 12: 19–36.
<https://doi.org/https://doi.org/10.25894/ldd161>.
- Thieberger, Nicholas. (2016) 2016. "Language Documentation." In *Languages of the Pacific Islands: Introductory Readings*, edited by H. Sato and J. Bradshaw. 2. Department of Linguistics.
<https://hdl.handle.net/11343/194673>.



Helpful language documentation texts



Gippert et al (2006)- fieldwork ethics, multimedia annotation, data processing



Bowern (2008)- funding applications, writing up data



Grenoble & Furbee-Losee (2010)- community collaboration, modern digital recording methods



Boerger et al (2016)- diverse topics, can carry out a documentary project from start to finish.

- Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel. 2006. *Essentials of Language Documentation*. Walter de Gruyter.
- Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. Palgrave Macmillan.
- Grenoble, Lenore, and Louanna Furbee-Losee. 2010. *Language Documentation Practice and Values*. Benjamins.
- Boerger, Brenda H., Sarah Moeller, Will Reiman, and Stephen Self. 2016. *Language and Culture Documentation Manual*. Leanpub. <https://leanpub.com/languageandculturedocumentationmanual>.



A brand new free
online textbook:
Archiving for the
Future: Simple Steps
for Archiving
Language
Documentation
Collections

- The Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin is announces the **second edition** of the free online textbook *Archiving for the Future: Simple Steps for Archiving Language Documentation Collections*.
- It is available at its new, dedicated website, <https://sites.utexas.edu/archivingforthefuture/>, as part of the University of Texas Libraries Open Books Pressbooks Publishing Network at <https://utexas.pressbooks.pub/archivingforthefuture/>.
- The target audience is anyone working to create materials for endangered or minority languages, especially those with no previous experience in archiving. The curriculum explains what to do before, during, and after data collection to facilitate long-term preservation of the data.
- The course material is based upon work supported by the National Science Foundation under Grant No. BCS-1653380 (Susan S. Kung and Anthony C. Woodbury, PIs; September 1, 2016, to August 31, 2020). The course is licensed under a Creative Commons Attribution-ShareAlike 4.0 International license, <https://creativecommons.org/licenses/by-sa/4.0/deed.en>.

Sullivant's organizational schema is the basis used for organizing the Kamasau language corpus guide.



Sullivant (2020) advocates for the depositor to help people find and reuse materials preserved in digital collections, instead of depending on the archivist to do it.



1) By providing adequate descriptive metadata, especially for audio-video materials that can not use text-based searches.



2) By preparing a guide that contextualizes recordings, summarizes their contents, and helps users find items in the collection.

Sullivant, Ryan. 2020. "Archival Description for Language Documentation Collections."
Language Documentation & Conservation 14: 520–78.
<http://hdl.handle.net/10125/24949>.



Sullivant's three priorities of metadata to include in a language corpus guide:

- First priority is information to discover a collection:
 - Collection identifier, the name and location of the repository
 - The extent and scope of the contents
 - Access restrictions that may be hidden in lower-level folders
- Second priority: recommended but not essential for initial archiving
 - Detailed contents list
 - Information about how items are arranged in the archive
- Third priority: optional but helpful information for users
 - Software needed to access materials
 - Physical access policies & times
 - Language data files expected to be added in the future.



Sullivant's
suggestions
applied to other
collections
giving insight to
the Kamasau
corpus guide:

- 1) Gawne (2018:213)- detailed formatting information for audio & video recordings of Syuba archived in two archives:
 - The Endangered Language Archive (ELAR) requested PFSX files that ELAN creates to be archived with the EAF files
 - Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADESIC) requested only EAF files
- Kamasau application: Knowing what SayMore derivative files could be archived in the SIL Archive.
- Gawne, Lauren. 2018. "A Guide to Syuba (Kagate) Language Documentation Corpus." *Language Documentation & Conservation* 12: 204–34. <http://hdl.handle.net/10125/24768>.



2) Caballero (2017:225) – the Choguita Rarámuri language guide included an example of extent of corpus and why:

- 130 hours of digital audio recordings
- 10 hours of video recordings
- The descriptive material included how and why materials were recorded based on a linguistic or a community –oriented agenda or both.

Kamasau application:

- Calculating extent of video and audio corpus already archived
 - Video recordings- 10 hrs 37 min
 - Audio recordings- 4.5 hrs (songs 3 hr 14 min, music-9 min, verb paradigms- 37 min, botany- 37 min)
- Deciding if the recordings were community driven or driven by linguistic goals.



Caballero, Gabriela. 2017. “Choguita Rarámuri (Tarahumara) Language Description and Documentation: A Guide to the Deposited Collection and Associated Materials.” *Language Documentation & Conservation* 11: 224–55. <http://hdl.handle.net/10125/24734>.

3) Franjeh (2019:94)- language of northern Ambrym, Vanuatu

- Handbrake program used to change video files to MP4s
- Those MP4 files were used in ELAN time-aligned files & to archive videos in the ELAR archive.
- Detailed metadata spreadsheet

Kamasau application:

- Proprietary MTS files used in the Cyberlink PowerDirector program were produced as MP4s after WAV and MTS files were synchronized, best audio file chosen, and subtitles added.
- The Kamasau project needs to archive detailed metadata spreadsheets

Franjeh, Michael. 2019. "The Languages of Northern Ambrym, Vanuatu: A Guide to the Deposited Materials in ELAR." *Language Documentation & Conservation* 13: 83–111.
<http://hdl.handle.net/10125/24849>.



4) Salffner (2015)

- Discussed detailed metadata about how the farm walk sessions were recorded & ethnobotanical data collected in the Ikaan yam gardens.
- Gave detailed information about how recordings were made.

Kamasau application:

- Ethnobotanical consultants were audio recorded on Zoom H2n Handy recorders.
- Sonya Paoli photographed parts of the plants in lieu of pressed botanical specimens.
- Canon EOS Rebel camera with 18-55mm lens for stills.
- Photos and audio recording later combined to make videos.



Salffner, Sophie. 2015. "A Guide to the Ikaan Language and Culture Documentation." *Language Documentation & Conservation* 9 (September): 237–67. <http://hdl.handle.net/10125/24639>.

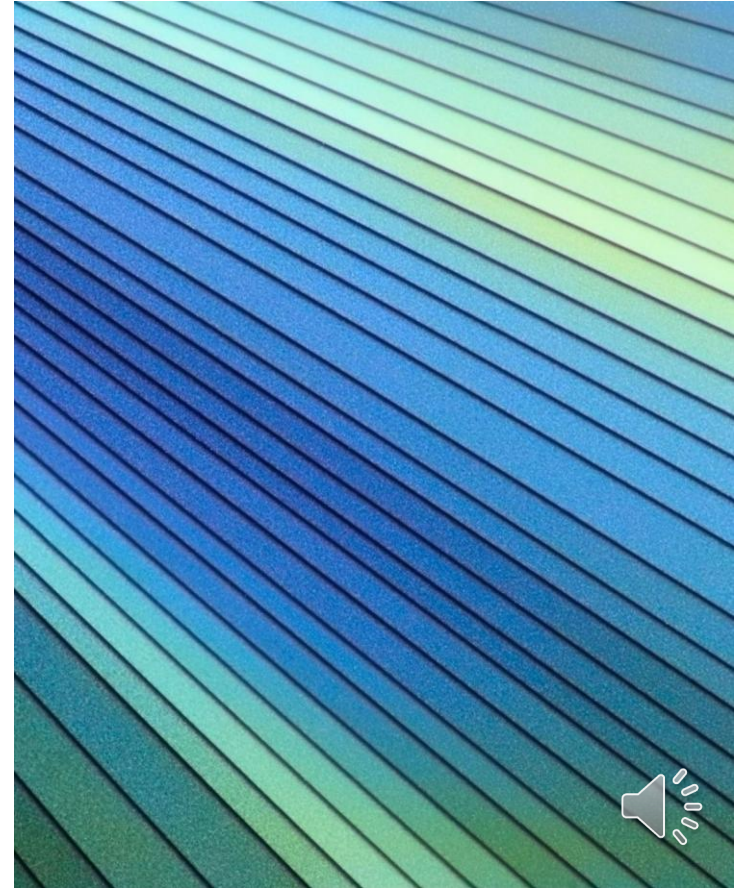
Summary

Following Sullivant's 3 priority level organizational schema, helped focus on what was important to include in the Kamasau language corpus guide.

First, adequate information to discover the collection.

Next, contents of the collection + background information about how data was collected.

Finally, software needed to access the corpus and plans for future additions to the corpus.



Filling the gaps

Working through Sullivant's guide, mistakes and missing items in the archive were found.

Examples:

- A book had been misnamed in archiving it.
- A missing song book was found in another collection.
- A paper written in 2022 had not been archived.
- The noun classes written only in the correction copy of the Triglot Dictionary need to be archived.
- The first 3 items were readily corrected; the latter will take time.



Checklist for reviewing a documentary corpus

- A review of the checklist in Boerger et al (2016) reminded the author of several additional missing items:
 - Metadata about how the legacy song contests were recorded and digitized from reel-to-reel recorders was missing.
 - Community members had not been taught how to access the SIL Archive.



Conclusion

Organizing data in the
Kamasau language corpus
guide into

Legacy Data and

2014 recordings and data,

clarified the contents of the
archive.



The process of creating the Kamasau corpus guide has two benefits:

- First, it helps users gain access to the Kamasau language corpus.
- Second, it gives ideas to others wanting to produce corpus guides in their own languages.

