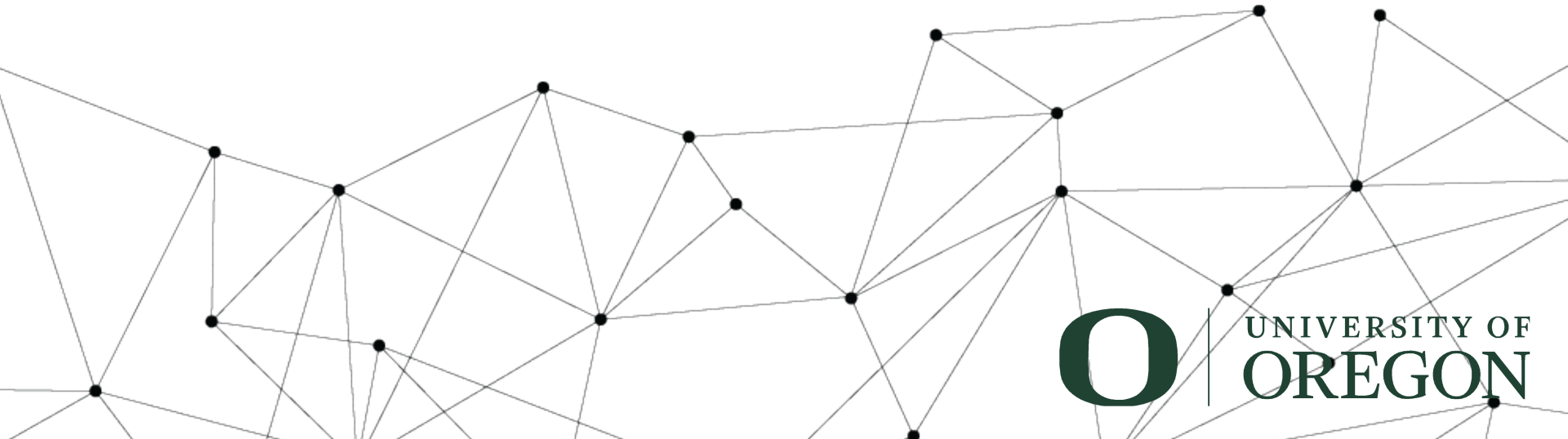
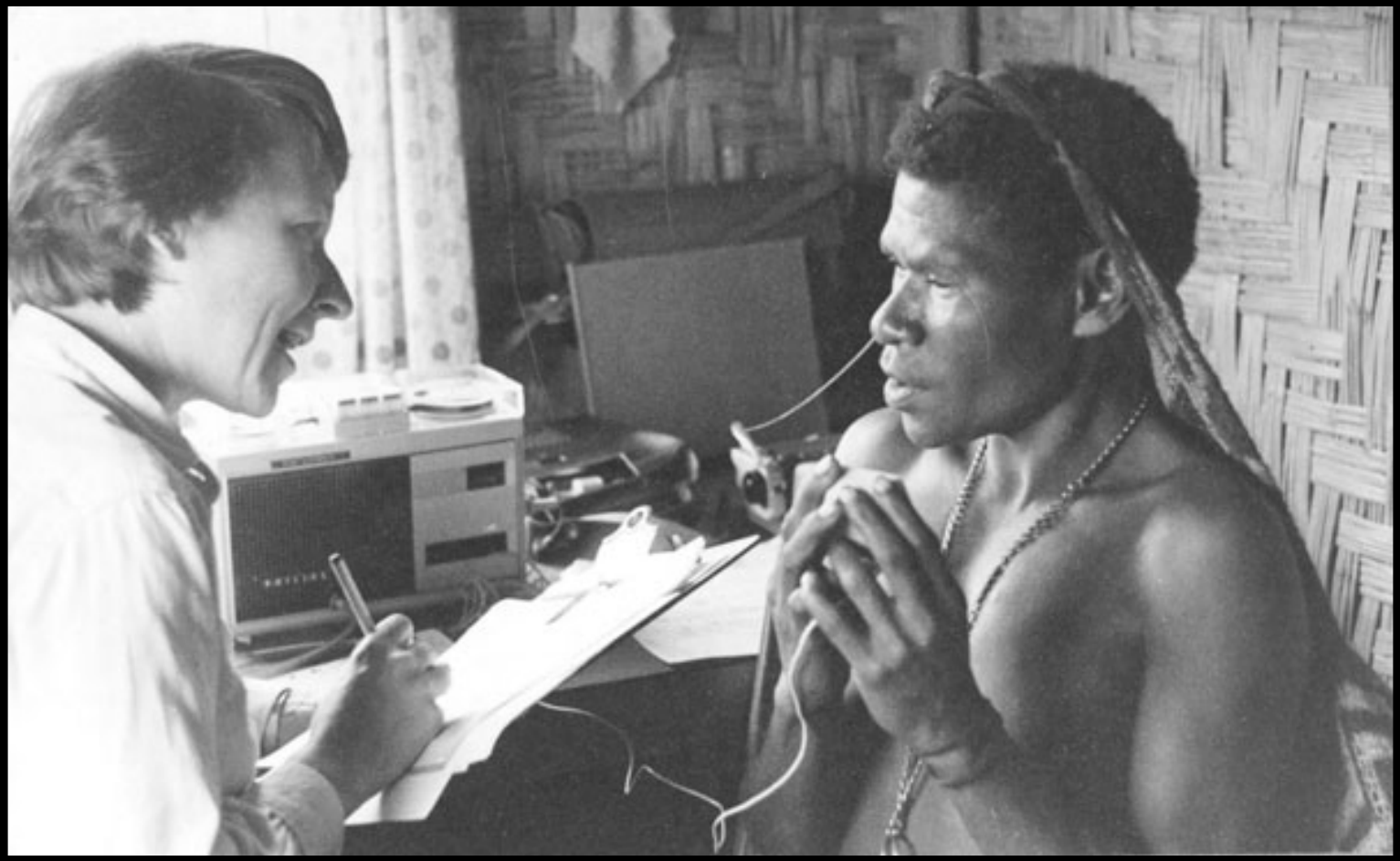
The digital notebook: a method for the rapid processing of elicited linguistic data

Richard T. Griscom, *University of Oregon*
Manuel A. Otero, *University of Oregon*

March 1st, 2019
ICLDC 6, University of Hawaii at Manoa



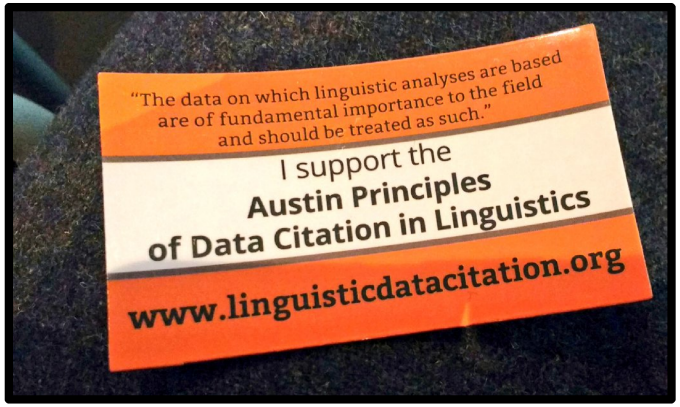
UNIVERSITY OF OREGON



FORCE11

The Future of Research Communications and e-Scholarship

Joint Declaration of Data Citation Principles

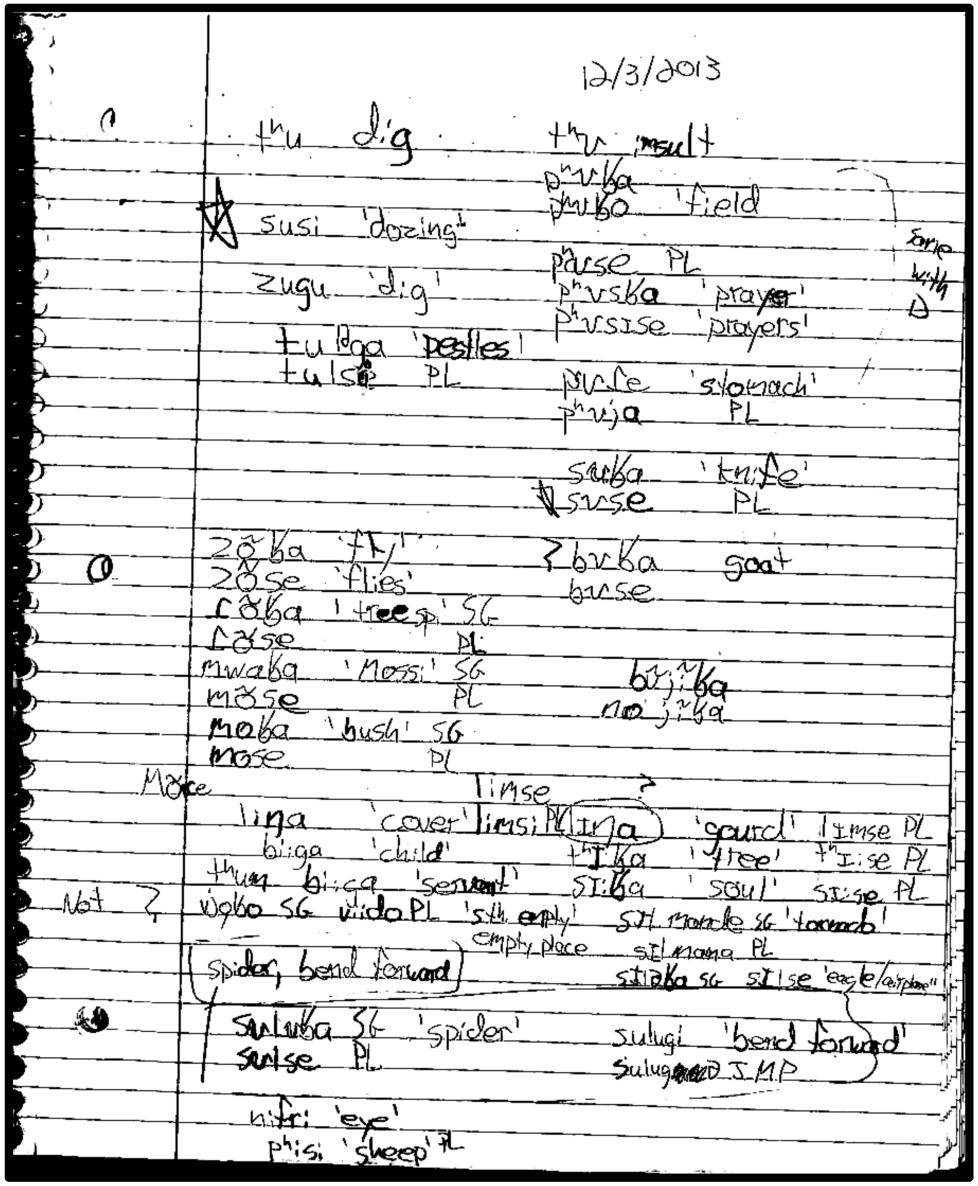


There are new expectations for data **citation** and **accessibility**

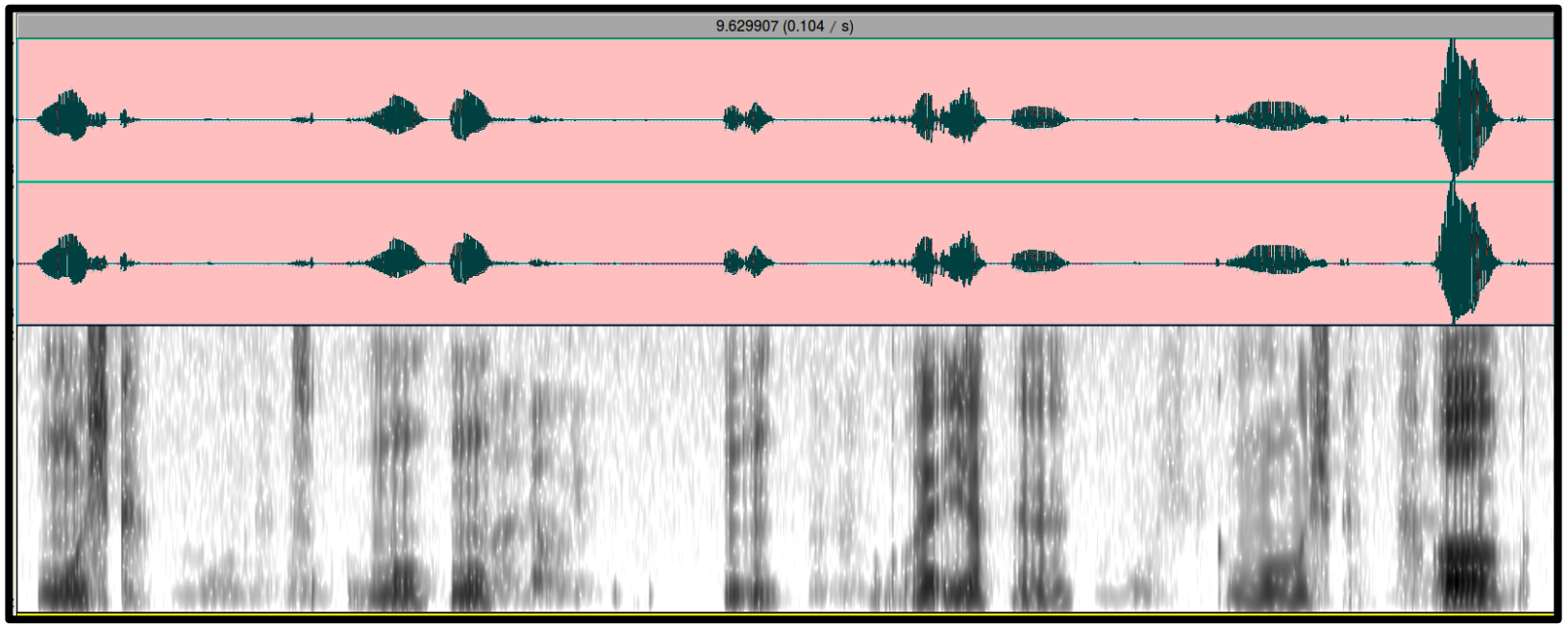
"...a lasting, multipurpose record..."

Himmelman (2006)

My field methods notebook

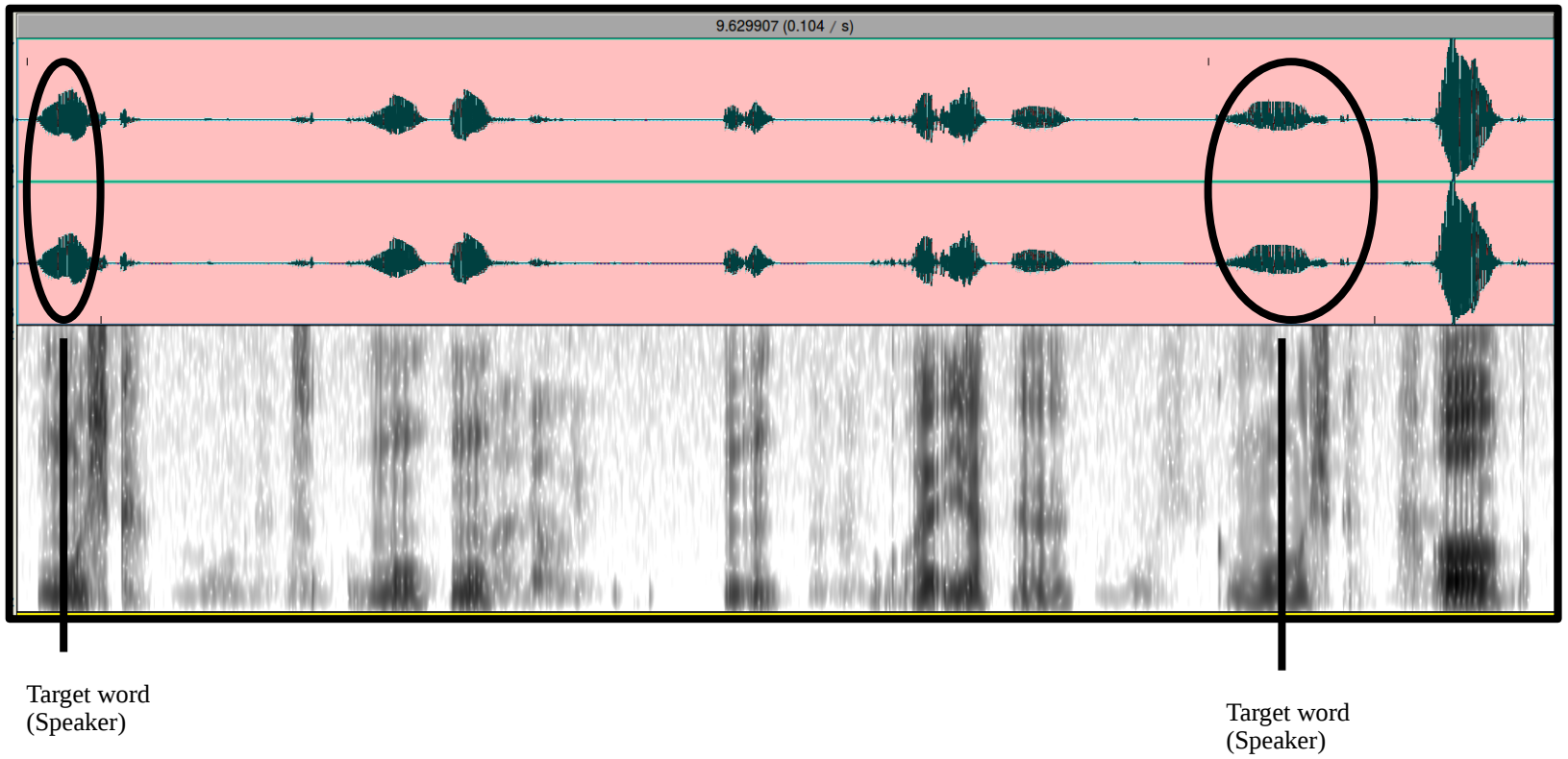


My first elicitation session

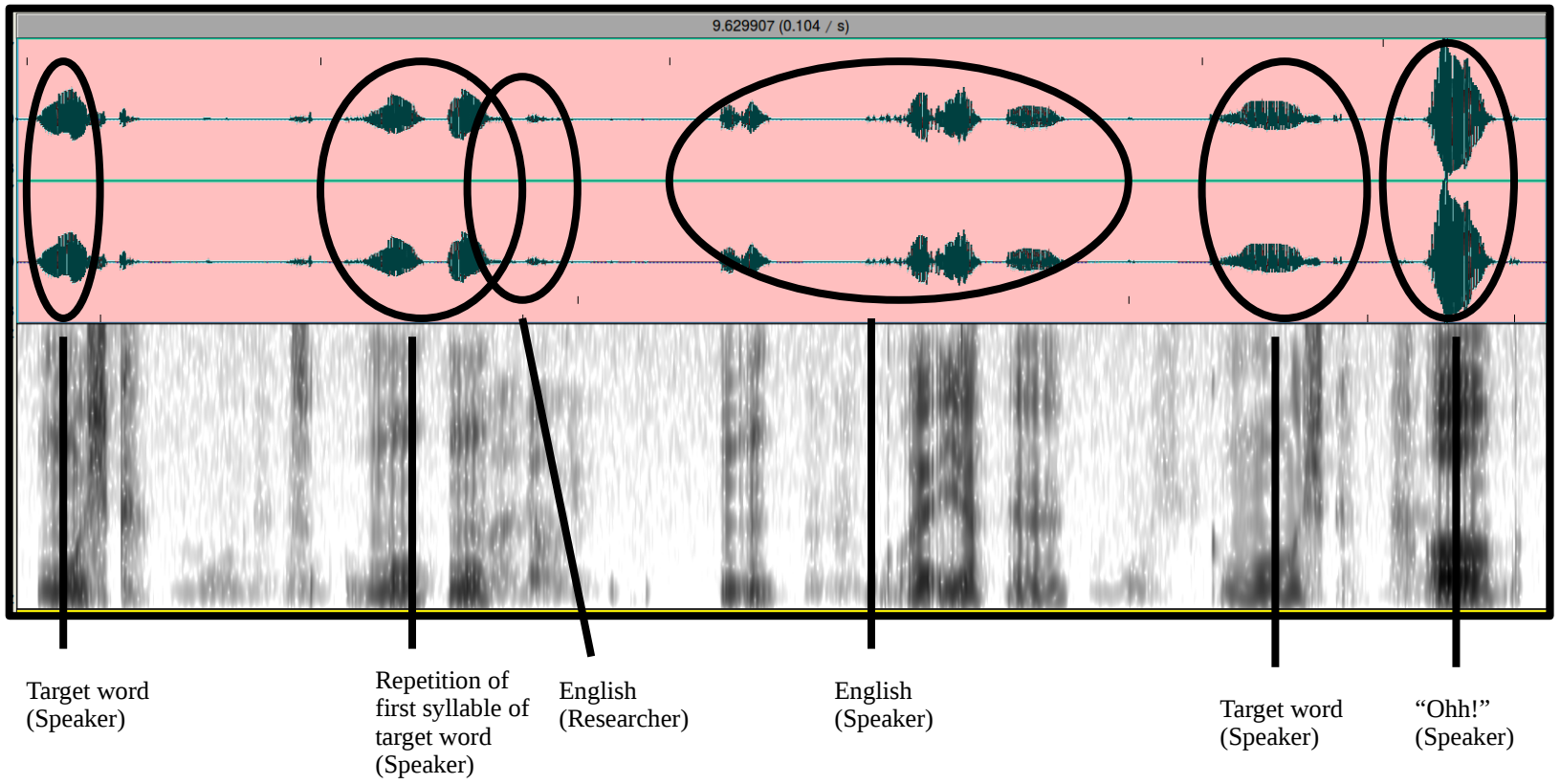


Asimjeeg Datooga speakers (Tanzania)

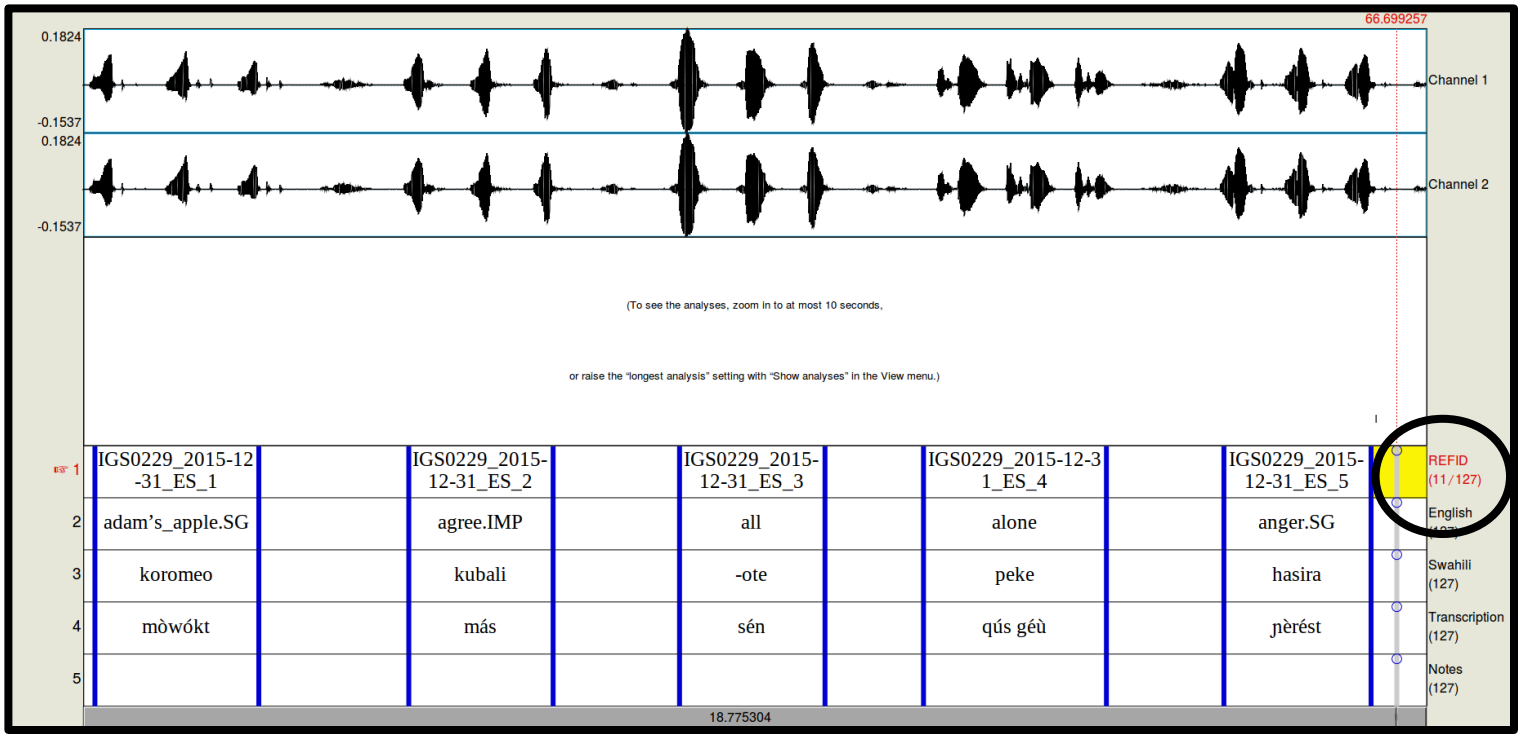
My first elicitation session



My first elicitation session

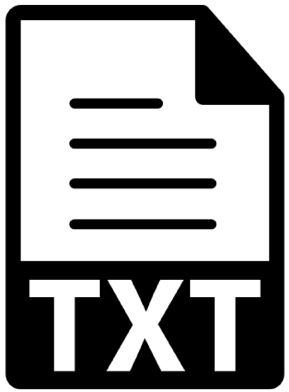


A goal for data citation and accessibility



1	IGS0229_2015-12-31_ES_1	adam's_apple.SG	koromeo	mòwókt		48.0833333333	50.4273333333
2	IGS0229_2015-12-31_ES_2	agree.IMP	kubali	más		52.5713333333	54.6353333333
3	IGS0229_2015-12-31_ES_3	all	-ote	sén		56.4433333333	58.5233333333
4	IGS0229_2015-12-31_ES_4	alone	peke	qús géù		59.9316431215	62.5473333333
5	IGS0229_2015-12-31_ES_5	anger.SG	hasira	ɲèrést		64.2273333333	66.3313333333

+



The Digital Notebook Method

- Provides nearly **instant access** to time-aligned elicited data
- **Scales up** for large data sets
- Produces **archive-ready** and **citable** documentation in three useful formats

Three main principles

Whenever possible...

- Start with **digital** text data
- **Plan** and **structure** recording sessions
- Use **automated processing** methods

Requirements

Trained speaker

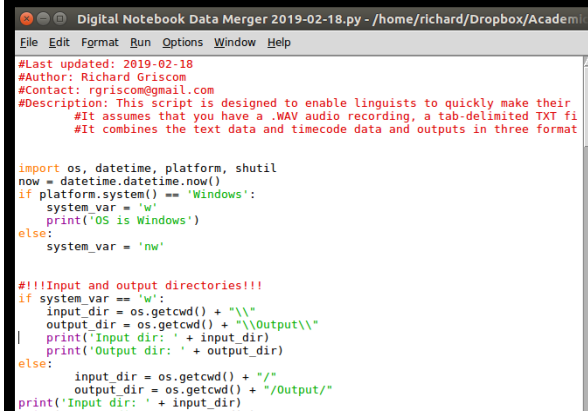
- Can consistently produce prompted elicited language, with repetitions

Recording equipment

- Audio recorder + headset microphone

Computer + software

- IPA input
- Data Merger program (optional)



Digital Notebook Data Merger 2019-02-18.py - /home/richard/Dropbox/Academi

File Edit Format Run Options Window Help

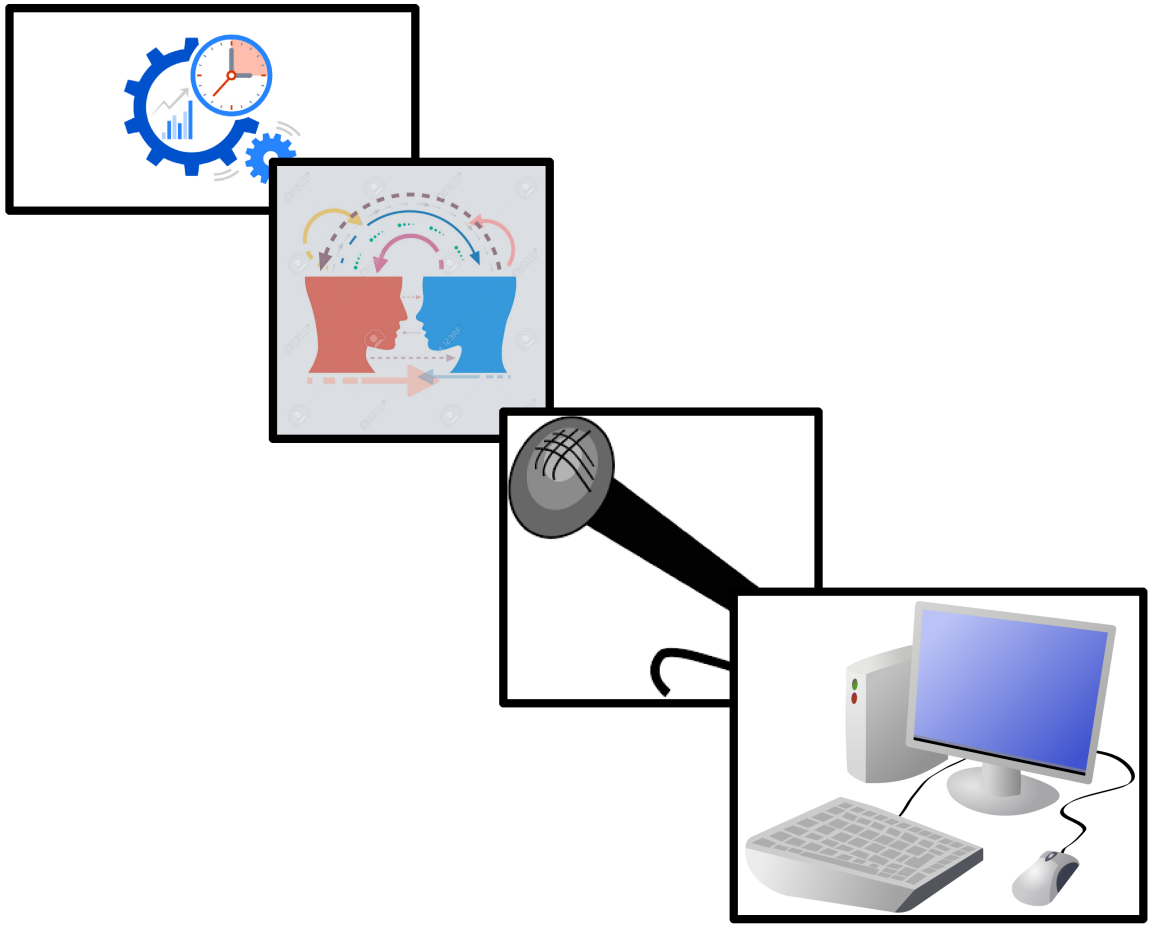
```
#Last updated: 2019-02-18
#Author: Richard Griscom
#Contact: rgriscom@gmail.com
#Description: This script is designed to enable linguists to quickly make their
        #It assumes that you have a .WAV audio recording, a tab-delimited TXT fi
        #It combines the text data and timecode data and outputs in three format


import os, datetime, platform, shutil
now = datetime.datetime.now()
if platform.system() == 'Windows':
    system_var = 'w'
    print('OS is Windows')
else:
    system_var = 'nw'


#!!!Input and output directories!!!
if system_var == 'w':
    input_dir = os.getcwd() + "\\"
    output_dir = os.getcwd() + "\\Output\\"
    print('Input dir: ' + input_dir)
    print('Output dir: ' + output_dir)
else:
        input_dir = os.getcwd() + "/"
        output_dir = os.getcwd() + "/Output/"
print('Input dir: ' + input_dir)
```

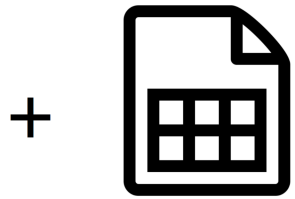
4 Stages of the Digital Notebook Method

1. **Preparation**
2. **Elicitation**
3. **Recording**
4. **Processing**



1. Preparation

Prepare your files for your session

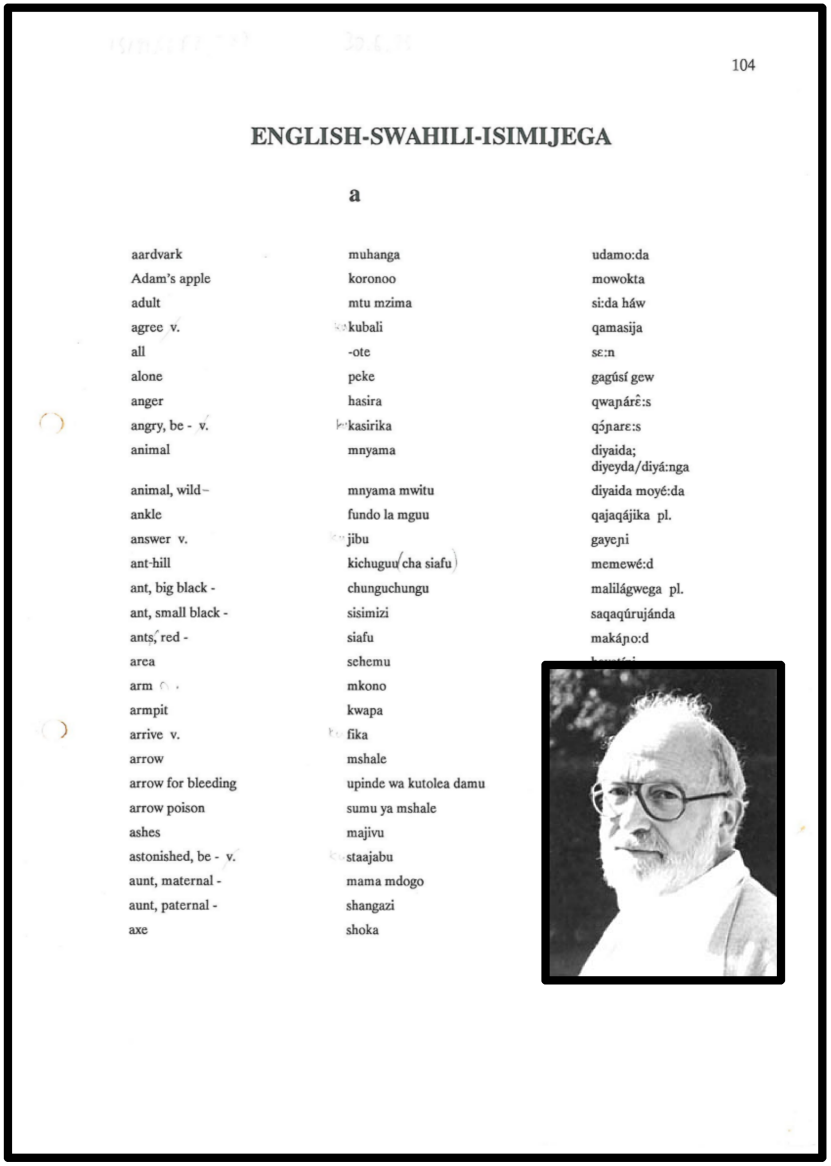


Elicitation page

1. Preparation

Plan your session

1	**Date:**	2015-12-31		
2	**Speaker:**	Eliya Shauri (ES)		
3	**Researcher:**	Richard T Griscom		
4	**Topic:**	Checking words from Rottland's Isimjeeg Dictionary		
5				
6	**English**	**Swahili**	**Transcription**	**Notes**
7	adam's_apple.SG	koromeo		
8	agree.IMP	kubali		
9	all	-ote		
10	alone	peke		
11	anger.SG	hasira		
12	be_angry.IMP	kasirika		
13	animal.SG	mnyama		
14	animal.PL	wanyama		
15	wild_animal.SG	mnyama mwitu		
16	ankle.PL	fundo la mguu		
17	answer	jibu		
18	ant_hill.SG	kichuguu cha siafu		



104

ENGLISH-SWAHILI-ISIMIJEGA

a

aardvark	muhanga	udamo:da
Adam's apple	koronoo	mowokta
adult	mtu mzima	si:da háw
agree v.	kubali	qamasija
all	-ote	sɛ:n
alone	peke	gagúsí gew
anger	hasira	qwaɲárê:s
angry, be - v.	kasirika	qɔ́ɲare:s
animal	mnyama	diyaida; diyeyda/diyá:nga
animal, wild –	mnyama mwitu	diyaida moyé:da
ankle	fundo la mguu	qajaqájika pl.
answer v.	jibu	gayeɲi
ant-hill	kichuguu cha siafu	memewé:d
ant, big black -	chunguchungu	malilágwega pl.
ant, small black -	sisimizi	saqaqúrujánda
ants, red -	siafu	makáɲo:d
area	shemu	
arm	mkono	
armpit	kwapa	
arrive v.	fika	
arrow	mshale	
arrow for bleeding	upinde wa kutolea damu	
arrow poison	sumu ya mshale	
ashes	majivu	
astonished, be - v.	staajabu	
aunt, maternal -	mama mdogo	
aunt, paternal -	shangazi	
axe	shoka	

2. Elicitation

Write in the notebook!

	A	B	C	D
1	**Date:**	2015-12-31		
2	**Speaker:**	Eliya Shauri (ES)		
3	**Researcher:**	Richard T Griscom		
4	**Topic:**	Checking words from Rottland's Isimjeeg Dictionary		
5				
6	**English**	**Swahili**	**Transcription**	**Notes**
7	adam's_apple.SG	koromeo	mòwókt	
8	agree.IMP	kubali	más	
9	all	-ote	sén	ATR?
10	alone	peke	qús géù	
11	anger.SG	hasira	ɲèrést	
12	be_angry.IMP	kasirika	ɲèrés	Was Rottland's entry an inflected verb?
13	animal.SG	mnyama	dìjànɲánd	
14	animal.PL	wanyama	dìjáŋg	
15	wild_animal.SG	mnyama mwitu	dìjànɲánd mòhéd	
16	ankle.PL	fundo la mguu	gìdg ség	
17	answer	jibu		Says there is no word for reply, only words for "say"
18	ant_hill.SG	kichuguu cha siafu	dìlgwàdʒánd	
19	big_black_ant.SG	chungchungu	màlìlàgwánd	
20	big_black_ant.PL	chungchungu	màlìlàgwég	
21	small_black_ant.SG	sisimizi	sàqàqùrdʒánd	
22	small_black_ant.PL	sisimizi	sàqàqúrg	
23	red_ant.PL	siafu	màkàɲód	Speaker not certain
24	area.SG	sehemu	héd	

2. Elicitation

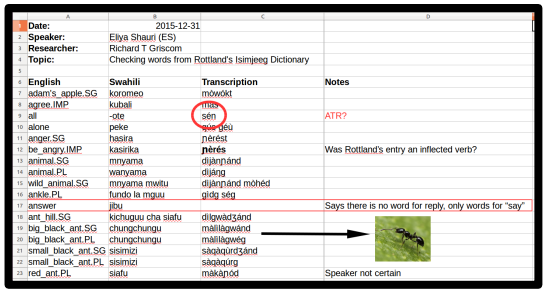
Write in the notebook!

- You can add images, text and cell formatting, etc. to your elicitation page

	A	B	C	D
1	**Date:**	2015-12-31		
2	**Speaker:**	Eliya Shauri (ES)		
3	**Researcher:**	Richard T Griscom		
4	**Topic:**	Checking words from Rottland's Isimjeeg Dictionary		
5				
6	**English**	**Swahili**	**Transcription**	**Notes**
7	adam's_apple.SG	koromeo	mòwókt	
8	agree.IMP	kubali	mas	
9	all	-ote	sén	ATR?
10	alone	peke	qúə gèù	
11	anger.SG	hasira	ɲèrést	
12	be_angry.IMP	kasirika	**ɲèrés**	Was Rottland's entry an inflected verb?
13	animal.SG	mnyama	dìjànɲánd	
14	animal.PL	wanyama	dìjáŋg	
15	wild_animal.SG	mnyama mwitu	dìjànɲánd mòhéd	
16	ankle.PL	fundo la mguu	gìdg ség	
17	answer	jibu		Says there is no word for reply, only words for "say"
18	ant_hill.SG	kichuguu cha siafu	dìlgwàdʒánd	
19	big_black_ant.SG	chungchungu	màlìlàgwánd	
20	big_black_ant.PL	chungchungu	màlìlàgwég	
21	small_black_ant.SG	sisimizi	sàqàqùrdʒánd	
22	small_black_ant.PL	sisimizi	sàqàqúrg	
23	red_ant.PL	siafu	màkàɲód	Speaker not certain

3. Recording

Prepare for the recording



Date:	2015-12-31		
Speaker:	Eliya Shauri (ES)		
Researcher:	Richard T Griscom		
Topic:	Checking words from Rottland's Isimjeeg Dictionary		
English	Swahili	Transcription	Notes
adam's_apple.SG	koromeo	mòwókt	
agree.IMP	kubali	más	
all	-ote	sén	ATR?
alone	peke	qús qéù	
anger.SG	hasira	ɲèrést	
be_angry.IMP	kasirika	ɲèrés	Was Rottland's entry an inflected verb?
animal.SG	mnyama	dìjàɲɲánd	
animal.PL	wanyama	dìjáŋg	
wild_animal.SG	mnyama mwitu	dìjàɲɲánd mòhéd	
ankle.PL	fundo la mguu	gìdg ség	
answer	jibu		Says there is no word for reply, only words for "say"
ant_hill.SG	kichuguu cha siafu	dílgwàdʒánd	
big_black_ant.SG	chungchungu	màlìlàgwánd	
big_black_ant.PL	chungchungu	màlìlàgwég	
small_black_ant.SG	sisimizi	sàqàqùrdʒánd	
small_black_ant.PL	sisimizi	sàqàqúrg	
red_ant.PL	siafu	màkàɲód	Speaker not certain

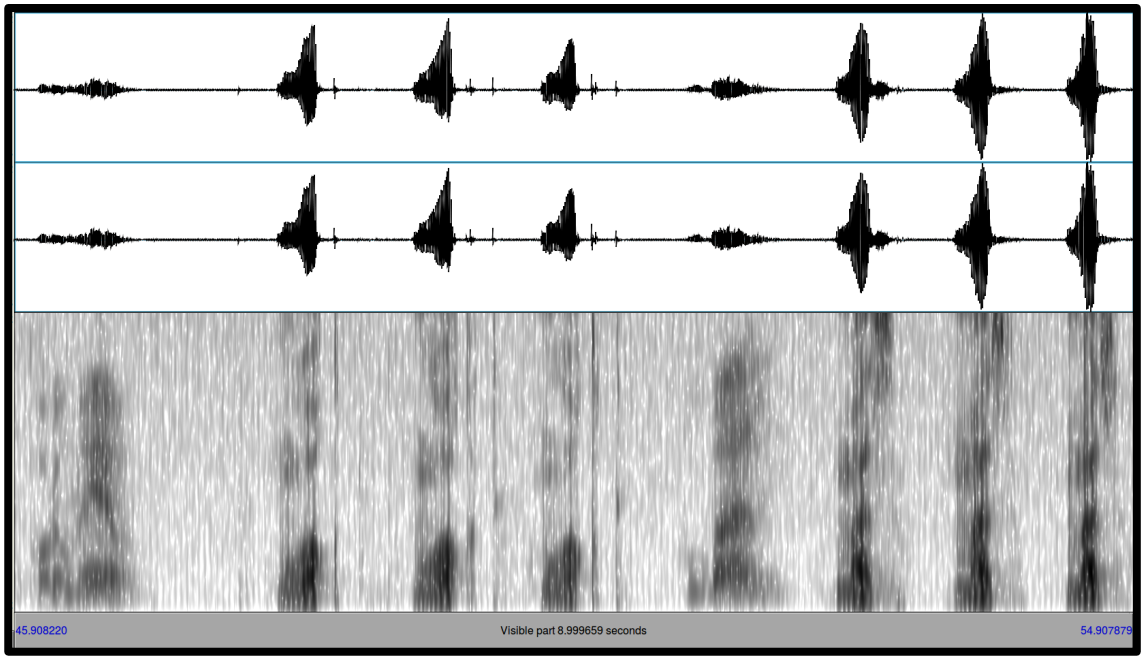
Elicitation page

	A	B	C	D
1	adam's_apple.SG	koromeo	mòwókt	
2	agree.IMP	kubali	más	
3	all	-ote	sén	ATR?
4	alone	peke	qús qéù	
5	anger.SG	hasira	ɲèrést	
6	be_angry.IMP	kasirika	ɲèrés	Was Rottland's entry an inflected verb?
7	animal.SG	mnyama	dìjàɲɲánd	
8	animal.PL	wanyama	dìjáŋg	
9	wild_animal.SG	mnyama mwitu	dìjàɲɲánd mòhéd	
10	ankle.PL	fundo la mguu	gìdg ség	
11	ant_hill.SG	kichuguu cha siafu	dílgwàdʒánd	
12	big_black_ant.SG	chungchungu	màlìlàgwánd	
13	big_black_ant.PL	chungchungu	màlìlàgwég	
14	small_black_ant.SG	sisimizi	sàqàqùrdʒánd	
15	small_black_ant.PL	sisimizi	sàqàqúrg	
16	red_ant.PL	siafu	màkàɲód	Speaker not certain

Recording page

3. Recording

Do the recording!



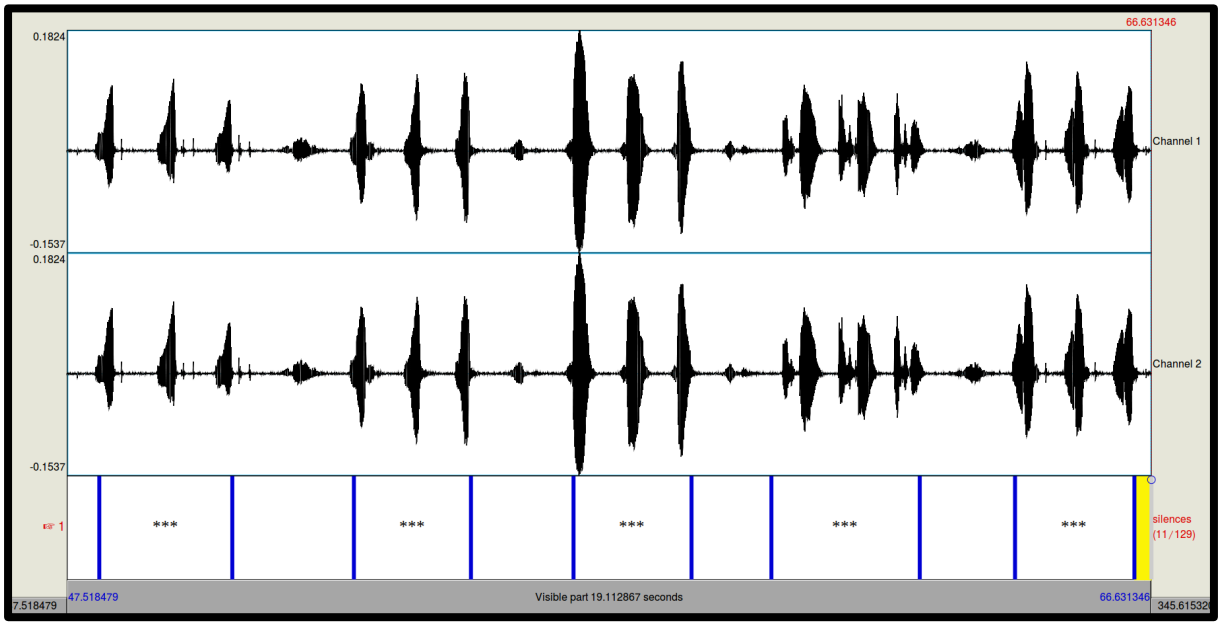
Structured recording with repetitions

4. Processing

Backup, backup, backup!

4. Processing

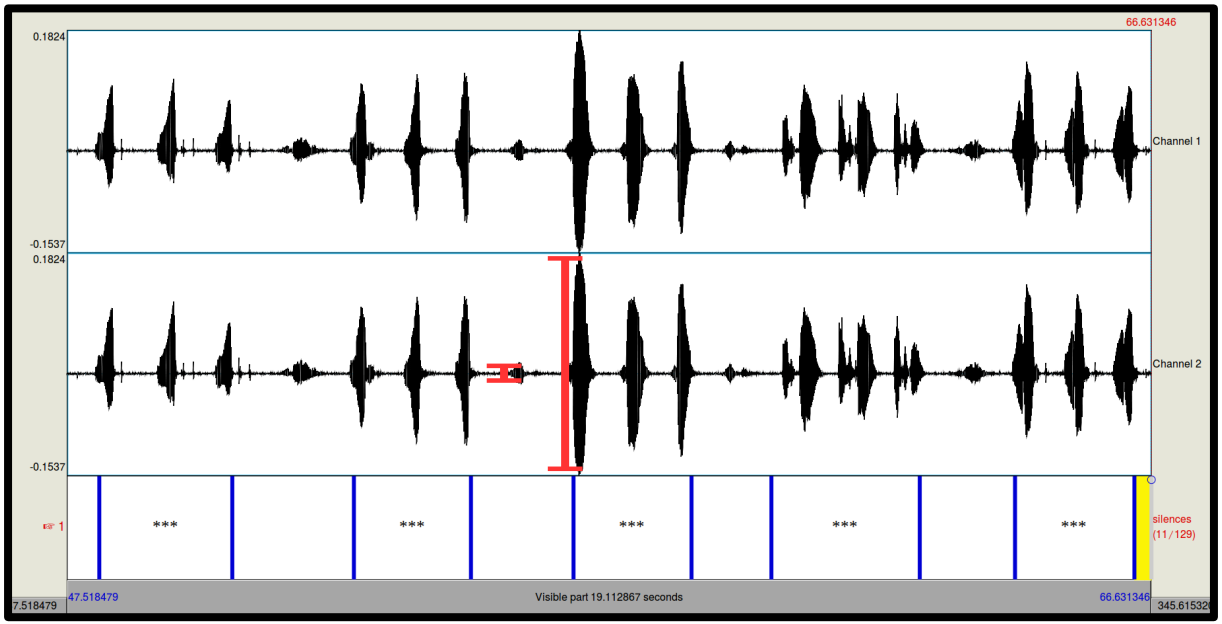
Use Praat to automatically segment the audio



Segmented audio with no text in Praat

4. Processing

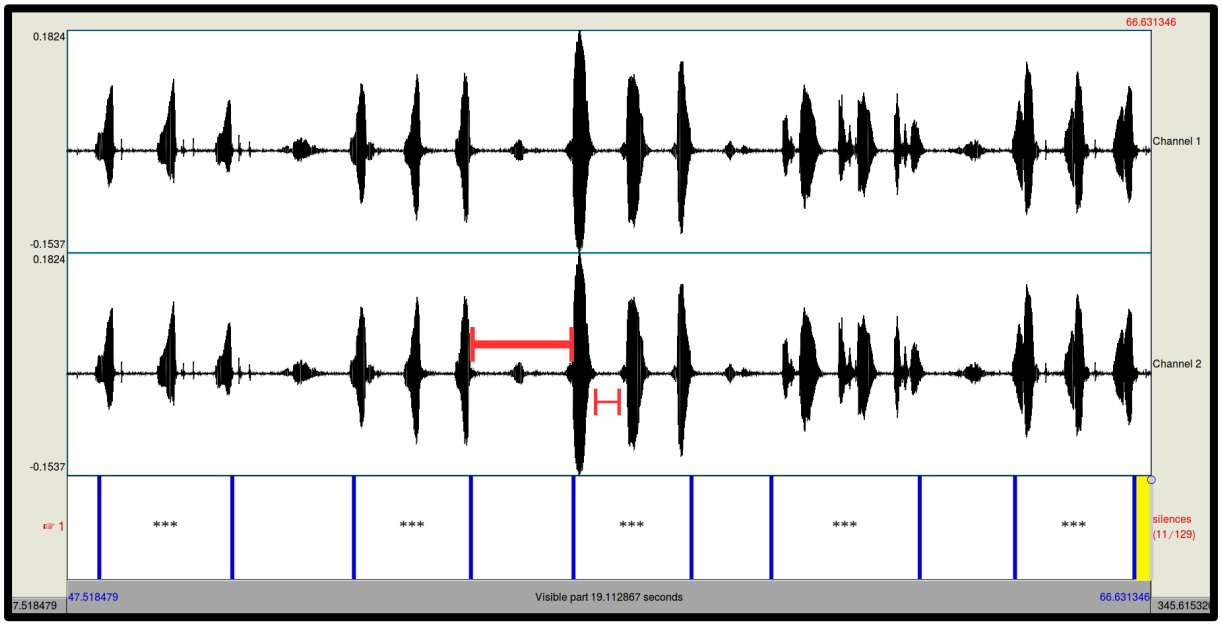
Use Praat to automatically segment the audio



Segmented audio with no text in Praat

4. Processing

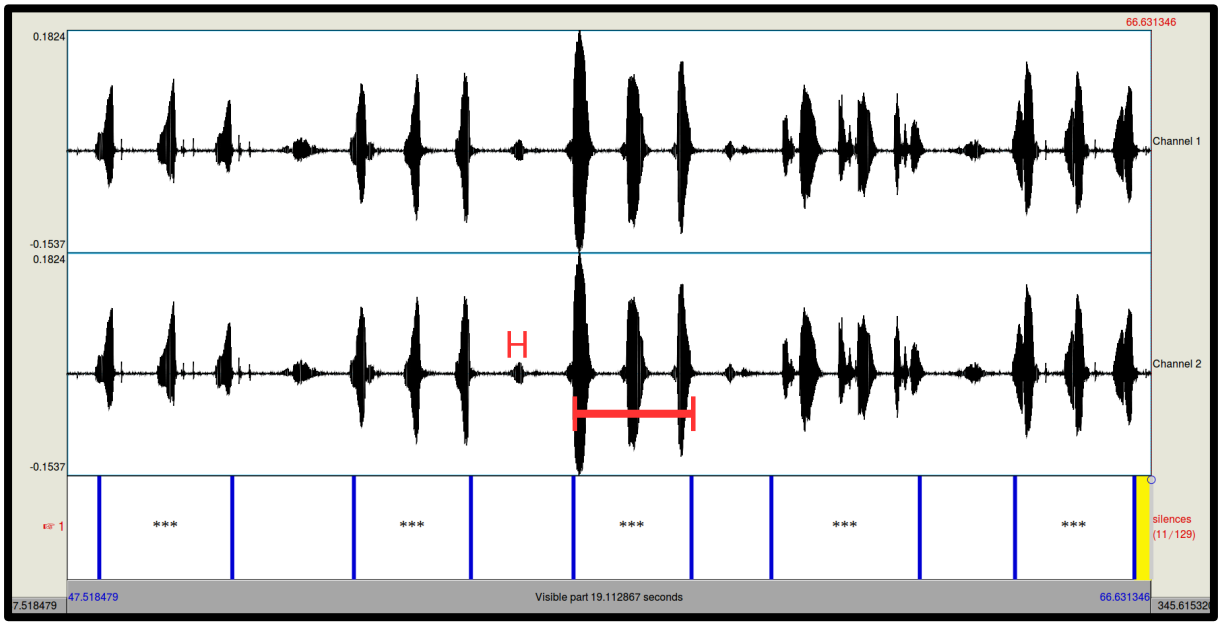
Use Praat to automatically segment the audio



Segmented audio with no text in Praat

4. Processing

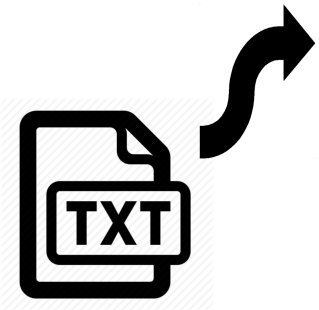
Use Praat to automatically segment the audio

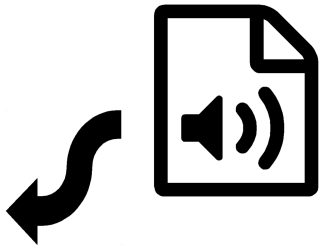


Segmented audio with no text in Praat

4. Processing

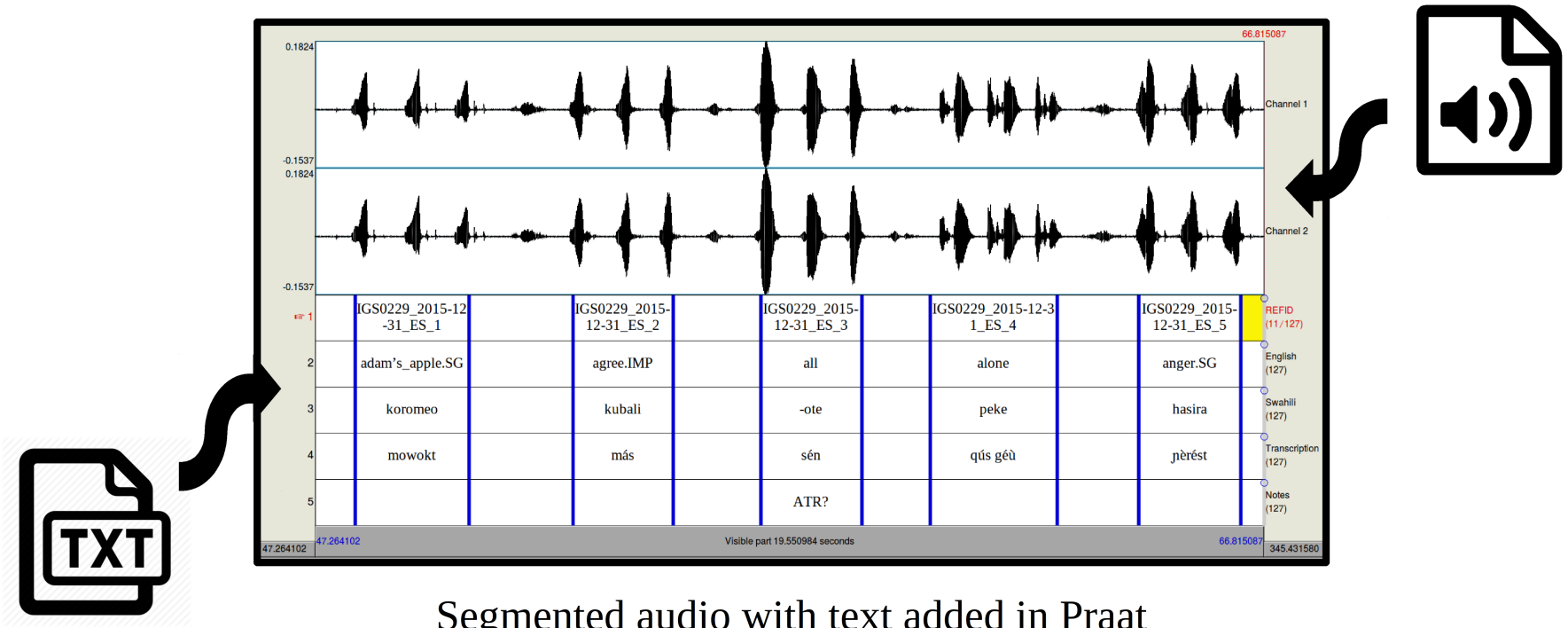
Use the Data Merger program to create time-aligned annotations





4. Processing

Use the Data Merger program to create time-aligned annotations



Segmented audio with text added in Praat

4. Processing

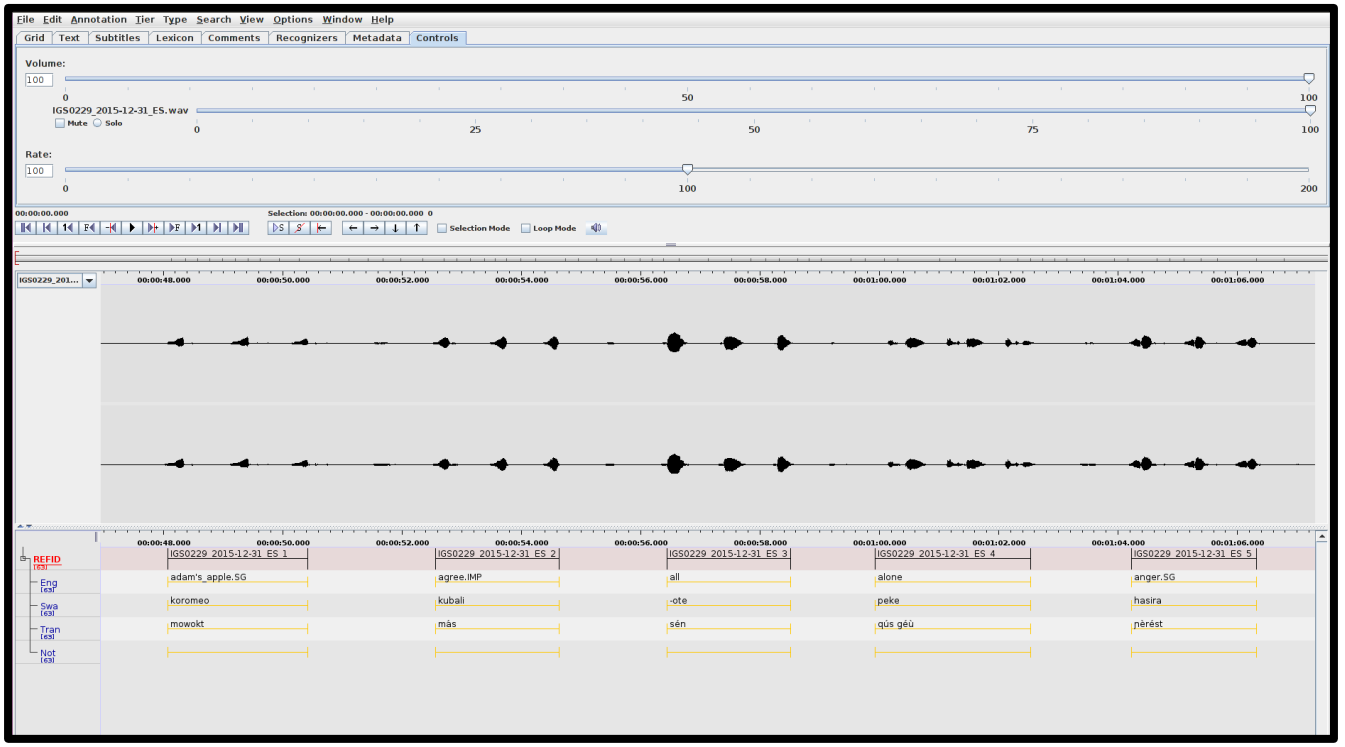
Use the Data Merger program to create time-aligned annotations

	Reference ID	English	Swahili	Transcription	Notes	Start time	End time
	A	B	C	D	E	F	G
1	IGS0229_2015-12-31_ES_1	adam's_apple.SG	koromeo	mowokt		48.0833333333	50.4273333333
2	IGS0229_2015-12-31_ES_2	agree.IMP	kubali	más		52.5713333333	54.6353333333
3	IGS0229_2015-12-31_ES_3	all	-ote	sén	ATR?	56.4433333333	58.5233333333
4	IGS0229_2015-12-31_ES_4	alone	peke	qús géù		59.9316431215	62.5473333333
5	IGS0229_2015-12-31_ES_5	anger.SG	hasira	nèrést		64.2273333333	66.3313333333

Tab-delimited text, opened in a spreadsheet application

4. Processing

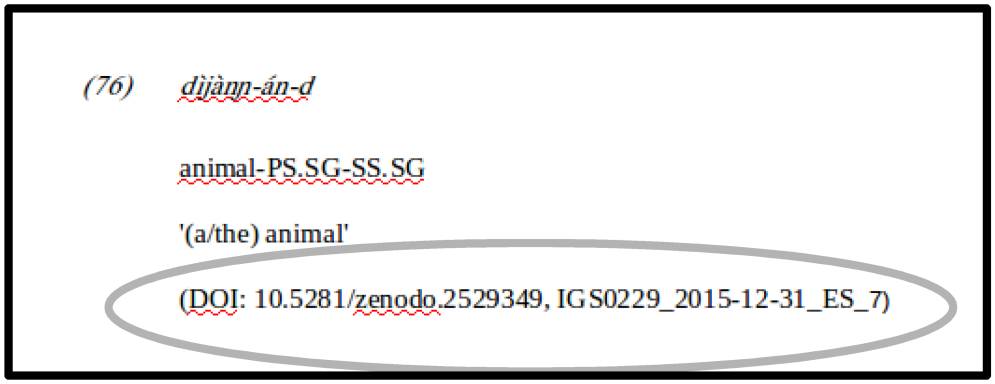
Use the Data Merger program to create time-aligned annotations



Time-aligned annotations in ELAN

Rejoice!

You have now successfully created accessible and citable data!



(76) *dìjànŋ-án-d*

animal-PS.SG-SS.SG

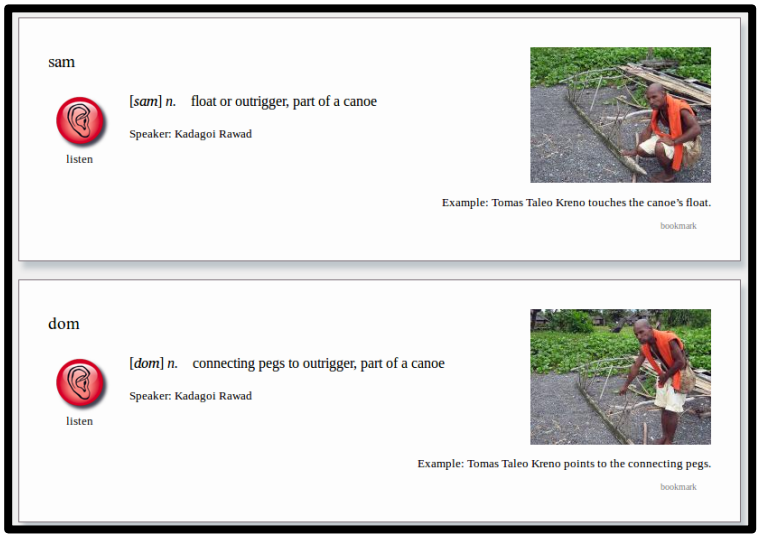
'(a/the) animal'

(DOI: 10.5281/zenodo.2529349, IGS0229_2015-12-31_ES_7)

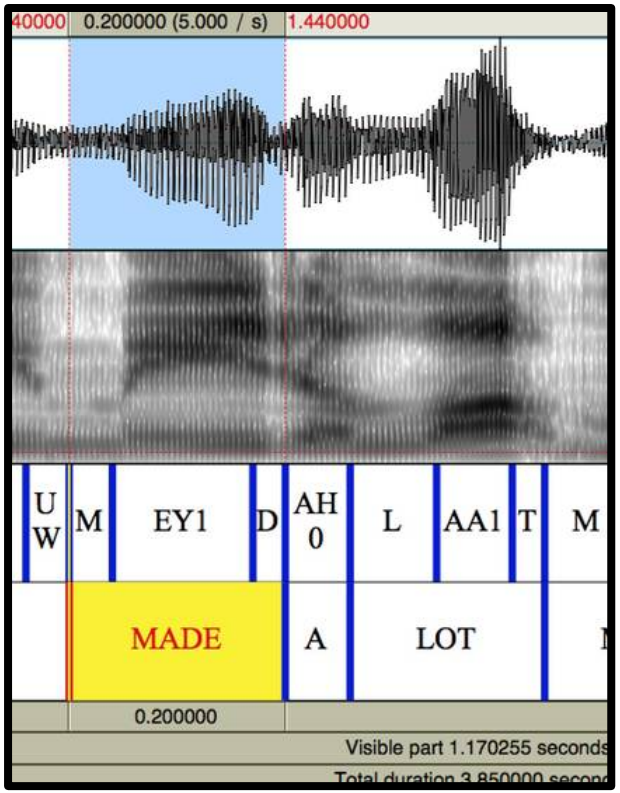
It’s also ready for archiving!

Applications

Talking Dictionary



Forced Alignment



Field methods corpus...

How does it work?

- Text data is **digital** from start to finish
- Recording sessions are **structured**
- Software is utilized for **automated processing**

The Digital Notebook Method

A good solution for **anyone** who plans to collect elicited data.

- **Instant access -** no more manual processing, saving countless hours of work.
- **Scalability -** no data set is too big.
- **Archive-ready** and **citable** data **-** improved accessibility and better science.

Resources

For learning the finer details of the method...

Wiki:
https://tinyurl.com/DigitalNotebookWiki

Video demonstration of processing stage:
https://youtu.be/NzCEfEZK4fw

Data Merger program:
https://tinyurl.com/DigitalNotebookDataMerger

Mahalo! / Thank you!

References

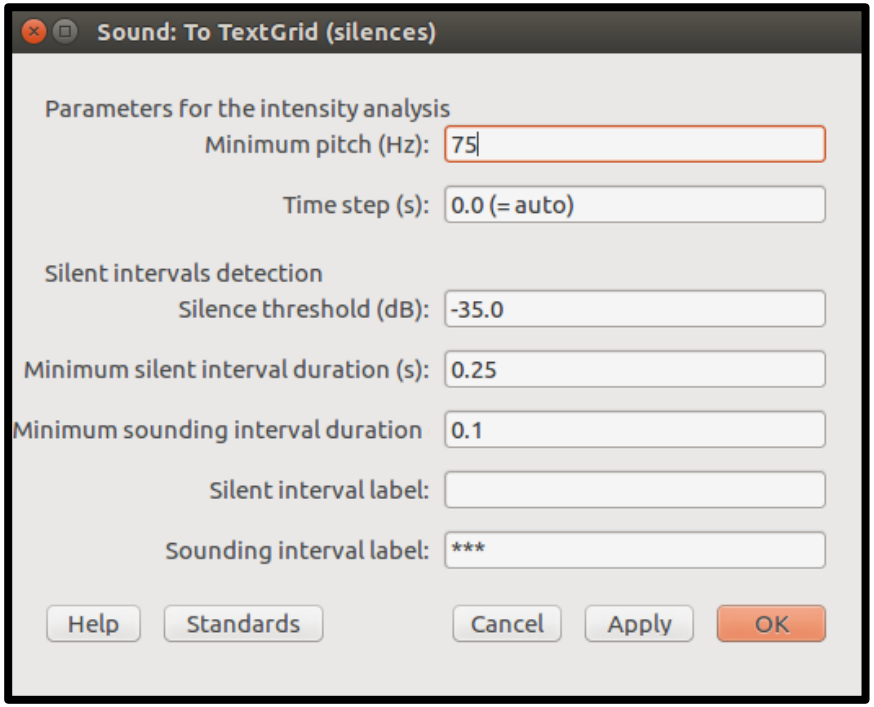
Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. 2018. The Austin Principles of Data Citation in Linguistics. Version 1.0. http://site.uit.no/linguisticsdatacitation/austinprinciples/ Accessed 2019-02-22

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

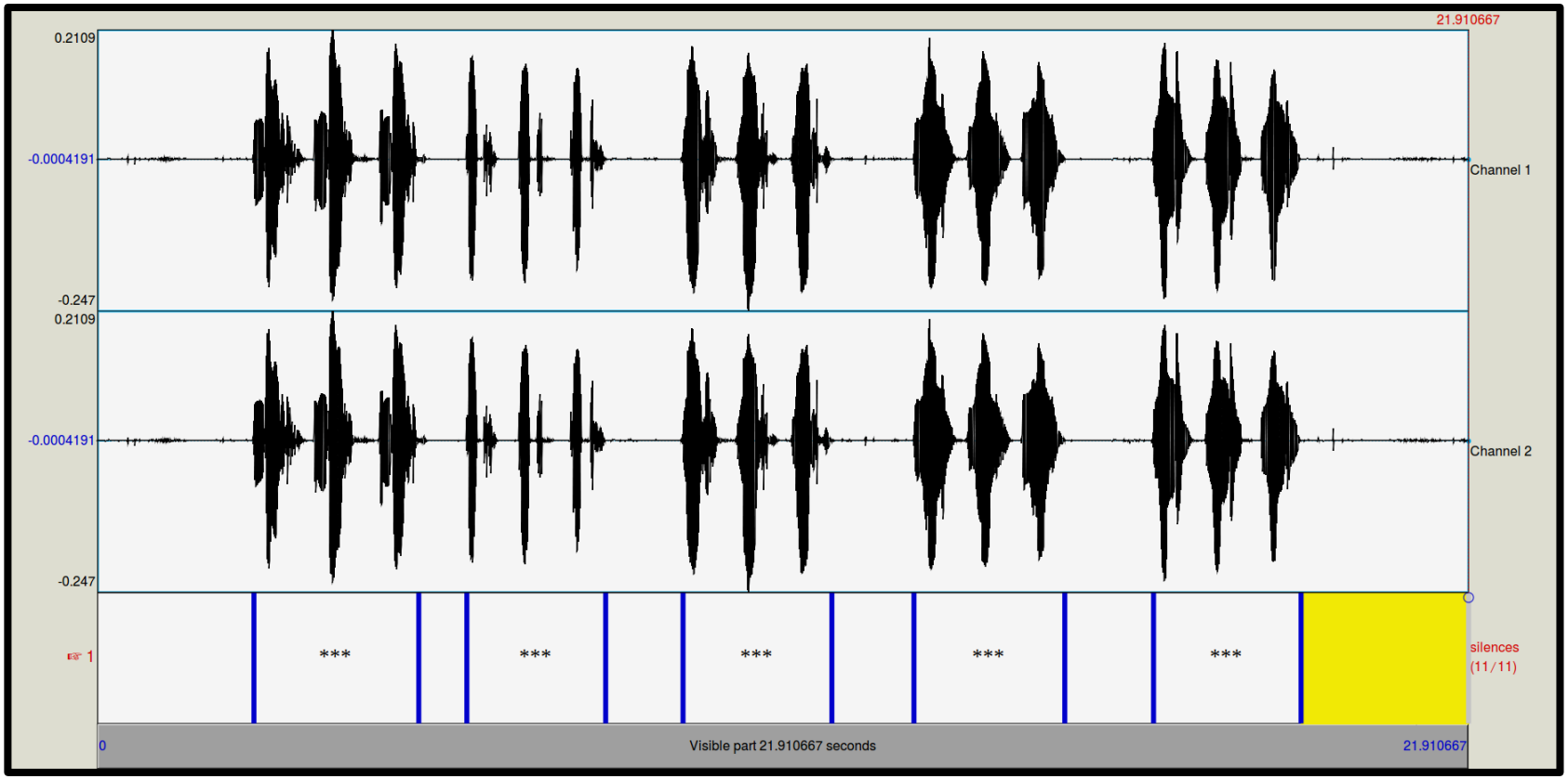
Griscom, Richard T. (2019) Documentation of Isimjeeg Datooga. https://elar.soas.ac.uk/Collection/MPI971096

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) Essentials of Language Documentation (Trends in Linguistics. Studies and Monographs, 178), 1-30. Berlin: Mouton de Gruyter.

Automatic segmentation in Praat

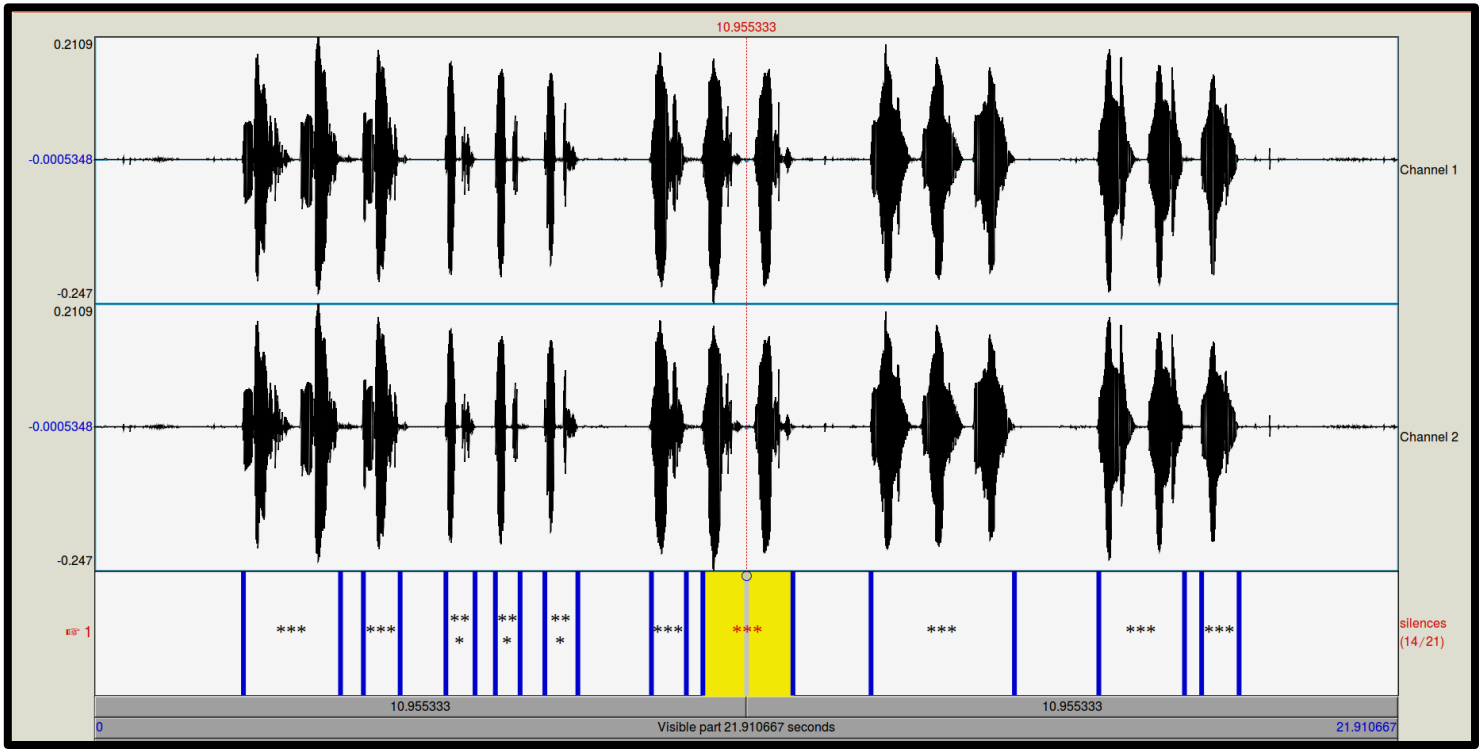


Automatic segmentation in Praat



What if the segmentation isn’t perfect?

- You can make any manual adjustments necessary to the segments before saving the TextGrid file.



Data Merger program

Currently....

If using Windows - There is a .EXE executable file that you can download from the Wiki. Put it in the same folder as the files you want to process.

If using Linux or Mac OS - Download Python + IDLE and the Data Merger python script. Put the script in the same folder as the files you want to process and run the script in IDLE.

Data Merger program

Digital Notebook Data Merger 2019-02-18.py - /home/richard/Dropbox/Academic

File Edit Format Run Options Window Help

```
#Last updated: 2019-01-05
#Author: Richard Griscom
#Contact: rgriscom@gmail.com
#Description: This script is designed to enable linguists to quickly make their
        #It assumes that you have a .WAV audio recording, a tab-delimited TXT fi
        #It combines the text data and timecode data and outputs in three format


import os, datetime, platform, shutil
now = datetime.datetime.now()
if platform.system() == 'Windows':
    system_var = 'w'
    print('OS is Windows')
else:
    system_var = 'nw'


#!!!Input and output directories!!!
if system_var == 'w':
    input_dir = os.getcwd() + "\\"
    output_dir = os.getcwd() + "\\Output\\"
    print('Input dir: ' + input_dir)
    print('Output dir: ' + output_dir)
else:
        input_dir = os.getcwd() + "/"
        output_dir = os.getcwd() + "/Output/"
        print('Input dir: ' + input_dir)
        print('Output dir: ' + output_dir)

dir_list = os.listdir(input_dir)
if "Output" in dir_list:
    try:
        shutil.rmtree(output_dir)
    except OSError as e:
        print ("Error: %s - %s." % (e.filename, e.strerror))
os.mkdir(output_dir)
total_columns = int(input('How many columns of text data? '))
counter = 1
column_names = []
while counter < (total_columns + 1):
```

Ln: 15 Col: 21

Data Merger program

- The script asks for the number of columns of text and the names for each column

- It is not limited to any number of columns!

```
How many columns of text data? 2
What do you want to label column #1? Translation
What do you want to label column #2? Transcription
['Translation', 'Transcription']
REFID   XMIN    XMAX    Translation     Transcription
CNNS_[sw]_2018-1-31_#1_1        2.4924761904761907      5.13247619047619        banana  ndizi
CNNS_[sw]_2018-1-31_#1_2        5.898190476190476       8.115333333333334       chair   kiti
CNNS_[sw]_2018-1-31_#1_3        9.349619047619047       11.726761904761904      table   meza
CNNS_[sw]_2018-1-31_#1_4        13.05247619047619       15.463904761904761      mountain        mlima
CNNS_[sw]_2018-1-31_#1_5        16.869619047619047      19.235333333333333      house   nyumba
Processing complete for: CNNS_[sw]_2018-1-31_#1.csv
Press Enter to continue...
>>>
```