Digital Notebook Method Tutorial

Richard Griscom, University of Oregon
rgriscom@gmail.com

Table of Contents

1 Introduction

2 Requirements

2.1 Trained speaker requirements

2.2 Hardware

2.3 Software

2.3.1 Installing LibreOffice

2.3.2 Installing Praat

2.3.3 Installing ELAN

2.3.4 Installing Python

3 Data management

3.1 Making a DMP

3.2 Making an archive plan

3.3 Practicing the workflow

3.4 Planning for version control

4 The Preparation Stage

4.1 Planning the session

4.2 File and folder preparation

4.2.1 Templates

5 The Elicitation Stage

5.1 Text data

5.2 Images

5.3 Metalinguistic data

5.4 Metadata

6 The Recording Stage

6.1 Training the speaker

6.2 Doing the recording

7 The Processing Stage

7.1 Backup

7.2 Segmenting the audio with Praat

7.3 Combining the text and audio data with the Data Merger script

8 The Archiving Stage

8.1 Preparing files for archiving

8.2 Depositing in the archive

9 Resources

10 References

Presentation: 20 minutes

1. Hook
2. Background

Overview of the digital notebook method:

- Designed for the rapid processing of elicited material from a single speaker
- Eliminates processing tasks that are dependent on the volume of data (e.g. segmenting audio, digitizing handwritten transcriptions)
- Prioritizes data management and archivability
- Gives fieldworkers more or less instant access to time-aligned corpora of their elicited data in the field and enables them to archive their data immediately after returning from the field

Required technology:

- Hardware
- Standard field recording equipment, ideally including a headset microphone
- A laptop for data processing
- Mobile device with bluetooth keyboard (only if it is not possible to also use the laptop during the elicitation session)
-
- Software
- IPA input software (KEYMAN, etc.)
- Spreadsheet software (LibreOffice/OpenOffice Calc, Excel, WPS Office)
- Praat
- ELAN

Data Management

- The digital notebook method requires a data management plan, including a clear organizational scheme, filename format, and metadata.
- File organization should include a primary folder for the project/language that contains the metadata for the project and a data folder. The files included in the data folder can be organized into folders named in the following format: [Date]_[Speaker Code]_[Session Number]
- Filenames should be consistent both across different files for an individual session and across different sessions. The filename should begin with project information (e.g. project

code/language code), followed by the date of the recording in ISO format, followed by session-specific information:

-

3. Method description

- 1. Preparation

Before the session

This section consists of the preparation before an elicitation session and includes the creation of spreadsheets designed to target specific linguistic phenomena.

1. **General templates:** pre-formatted spreadsheet templates are obtained or created for various types of elicitation sessions (word-list sessions, paradigm sessions, etc.), SHOW SOME TEMPLATES
2. **Session folder:** A new folder is created for the session and named in the format [Date]_[Speaker Code]_[Session Number]
3. **Session-specific digital notebook spreadsheet:** A template is modified accordingly to meet the goals of the upcoming elicitation session (i.e. writing translations in a word list, creating paradigms, etc.). The file is saved in the session folder with a filename in the standard format for the project with "_NB" inserted before the extension.

- 2. Notebook

During the elicitation session

This section includes two separate phases of the elicitation session: (1) the digital notebook phase, during which the fieldworker works with the consultant to create transcriptions and add metadata into the digital notebook spreadsheet, and (2) the recording phase, when the fieldworker creates a recording spreadsheet and records the consultant producing each item in the recording spreadsheet in the same order in which they appear.

Digital notebook phase

4. **Transcription data entry:** Transcriptions and metadata are entered into the digital notebook spreadsheet during this exploratory portion of the elicitation session. Any supplemental handwritten notes/drawings are made in a physical notebook, to be digitally captured later and with specific references to the entries in the digital notebook. This phase can optionally be recorded, ideally with an omnidirectional or stereo microphone.
5. **Save digital notebook spreadsheet:** The digital notebook spreadsheet should be saved exported into the PDF-A format with the same filename, and the file should be opened and checked before continuing.

6. **Elicitation-only spreadsheet:** A second spreadsheet including either all or a subset of the items elicited during the exploratory phase is created for the purpose of guiding the elicitation-only phase, with the same filename except that "_REC" is substituted for "_NB".

- 3. Recording

Data recording phase

7. **Data recording session:** a recording session is done specifically for the data that has been entered into the recording spreadsheet, in the same order, with repetitions and timing designed specifically with automated audio segmentation in mind. The speaker should be trained in producing repetitions of elicited items prior to the initiation of recording. A headset microphone is ideal for this recording.

8. **Save recording spreadsheet in processing folder:** The spreadsheet should be saved, but <u>not</u> exported to PDF-A.

- 4. Processing

  9. **Save and backup audio-visual recordings –** Recordings should be backed up on multiple devices and optionally cloud storage.

  10. **Automated audio segmentation:** Praat is used to automatically segment the audio recording.

  10.1 Open the Praat audio analysis software

  10.2 Open → Read From File… (Select the recording)

  (*Long Sound Files do not work!)

  10.3 Annotate → To TextGrid (silences)…

  10.4 Use the following settings:

  Minimum Pitch: 70
  Time Step: 0.0
  Silence Threshold: -35.0
  Minimum silent interval duration: 0.25
  Minimum sounding interval duration: 0.1
  Silent interval label:
  Sounding interval label: ***

  10.5 File → Save TextGrid as text file… (save with same filename as recording notebook)

4. Resources
5. Future
6. References

Handout:
1 paragraph summary of the method
Description of each stage
TinyURL links to resources

Website:
1 paragraph summary of the method
Description of each stage
Links to resources
Link to Youtube video

Youtube video #1:
-Intro to what the method is
-Demonstration of using the method

Youtube video #2 (after conference):
-Video of presentation

The problem:
-Scaling vs. citation
-Language documentation requires the collection of large amounts of data
-Developing data citation and archiving standards require linguists to make their data structured and accessible

We want to document and preserve languages
But it turns out that "doing documentation right" means collecting so much data

HOOK:

- Imagine you are sitting with a speaker of an endangered language, and you are writing notes on the words they are saying in your notebook and you are capturing their voice with an audio recorder. You complete the session and thank the speaker, and they go on their way. Now imagine that, after a few minutes alone, you suddenly realize that you had encountered a particular word from the session before, and miraculously within seconds you have managed to search all of the notes in all of your notebooks for every instance of that word, and you are able to hear audio recordings that correspond to each one. Today, I will show how this scenario isn't just imaginary, but can actually be reality.

-
- if, a few minutes after writing language data in your notebook, you could instantly search all of your notes and call up audio or video recordings that correspond to each piece of data.

-
- Language is complex, and our goal as documentarians is to make languages accessible, both for speakers and for non-speakers, and both for those living today and for those who have yet to come. (wide definition of accessibility)
- The resources available to language documentarians have always constrained their ability to make language data accessible.
- In the past, they might have only had a notebook and a pen, or perhaps a large and cumbersome reel-to-reel audio recorder. Later came cassette recorders and microphones, but recordings and notes were still analogue, which meant that accessibility was largely restricted to printed materials.
- In the past twenty years, there has been a digital revolution. Today, most language documentarians have access to digital audio recorders, computers, mobile technology, and the internet. All of these technologies give us the ability to share language data with anyone from nearly all corners of the globe.
- So why is our data still not accessible? Two main misconceptions: it's not easy (the "I'll collect data first and worry about making it accessible later" mindset), and it's not beneficial for the language documentarian
- This presentation aims to disprove both of these misconceptions by introducing one specific method for making data accessible that is both easy and instantly beneficial for language documentarians working on their own.
-
- It is called the digital notebook method, and it is designed to recreate the experience of writing in a physical notebook as closely as possible while simultaneously making the data in that notebook and all of your other notebooks instantly searchable and archiveable.
-
- If you work on documenting a language, this method can save you hours and hours of unnecessary work!
- The digital notebook method is designed specifically for elicited speech from a single speaker. This could include word lists, grammatical paradigms and constructions, and prompted clauses or short phrases. Elicited speech is often important at early stages of language documentation, when a language is not very well understood by documentarians.

Stages

-
- Our ability to make that data accessible, was limited. If we wanted to find a single word in a story, we would have to manually scan through pages and pages of a book, or listen to hours of recordings.

- Today, however, language documentarians have access to a vast array of digital recording and computational technology that have the power to make languages accessible in ways that were never thought possible. So how do we do it?

- This presentation focuses on making one kind of language data instantly accessible and archiveable.
-
-

Audio recordings

I'm going to share a method with you that enables you too make your

The solution:
-Remove any unnecessary steps
-Automate as much as possible
-Work backwards from desired outcome

Related talks at the conference:
Workshop:
Accelerating the analysis of your audio recordings with Untrained Forced Speech Alignment (Rolando Coto-Solano, Sally Akevai (Ake) Nicholas, Samantha Wray, and Tyler Peterson

Taking aim at the 'transcription bottleneck': Integrating speech technology into language documentation and conservation
Christopher Cox • Carleton University
Gilles Boulianne
With the widespread adoption of digital recording techniques, language documentation programs now ofen produce more audiovisual materials than can be annotated by hand. We demonstrate how state-of-the-art automatic segmentation, speaker diarization, and language identification methods from computational linguistics can be integrated into documentary workflows to help address this 'transcription botleneck'

New goals for software tools in language documentation
Kavon Hooshiar • University of Hawai'i at Mānoa

Sofware tools for language documentation, while successful at providing core functionalities, fall short of their users' needs due to insufcient interoperability across tools, failure to prioritize archiving and collaboration, etc. I propose foci for tool development beyond iterative improvement of existing tools, mainly the creation of a standardized data structure.

Talk: Linguistic and metalinguistic training to support use of audio 'chunks' in language revitalisation

Archival audio documentation can play an important role in language revitalisation contexts with few or no fluent speakers. However this data must be made available in a form that meets the needs of language teachers and learners. This paper suggests one approach in the context of Mangarrayi people at Jilkminggan

The ongoing Challenge of connecting speakers to archival language records

Language archives provide an excellent service for structuring collections and for making them accessible. Connecting speakers with archival sources scatered in diferent locations, only available via the internet, is addressed by our use of local subcollection—files with their descriptions exported from the collection and delivered on local wifi transmiters.

Organizing Linguistic Data for Language Revitalization: The Konkow Maidu Website and Database

Using archival materials collected fify years ago, three NEH grant team members have created a website to help revitalize the Konkow Maidu language of Central California. A linguist, web master and curriculum developer are combining three disciplines to put these contemporary, applicable tools in the hands of Konkow language learners.

Documentation does not Revitalize: Conflicting goals and ethical dilemmas

Despite persistent beliefs, documentation does not, per se, revitalize. These are two very diferent endeavors that target very diferent objectives. Even in cases where documentation includes collaborative practices with the community, benefits follow for the field of Linguistics but not for the vitality of linguistic

practices within the speaking community.

Workshop
TRANSCRIPTION ACCELERATION FOR LANGUAGE DOCUMENTATION WITH ELPIS
In this workshop we will use Elpis, an open source speech to text system, to train language models and obtain automated
first-pass transcriptions for languages with low quantities of data. This workshop is suitable for linguists and language
workers; no machine learning experience is required.

Training communities in documentation and technology: A model for the future
Documentation Training Center (LDTC) ofer a model for conducting a language documentation training workshop. Discussion centers around its strengths and limitations, its influence on the community, and the reproducibility of such a model.

A problem shared is a problem solved! - managing data the Open Source way
Robert Forkel • Max Planck Institute for the Science of Human History
The parallels between sofware development and management and data curation provide for a lot of potential re-use of tools and practices. This presentation explores this potential with a focus on unexpected opportunities like using a sofware package manager as client sofware to access language data in archives.

Long-Distance Fieldwork Online: Case Studies in China and Pakistan
The spread of internet access, even in remote areas, is erasing barriers in fieldwork that previously limited or even curtailed documentation projects. In this talk, the benefits and disadvantages of these methods are explored, drawing from the author’s experience as a US-based linguist with ongoing projects in China and Pakistan.