



Language Documentation @Field Methods

Intro

Brenda H. Boerger & Matthew Lee



**Who am I?
(And should you listen to me?)**

Matthew R. Lee



- MA in Applied Linguistics @ DIU 2022
 - BA in Integrated Science and Technology & BA in Philosophy and Religion @ James Madison University 2006
-
- Adjunct Faculty @ DIU since 2022
 - Field Data Management Instructor @ GIAL 2013
 - Instructor @ Cameroon Baptist Theological Seminary 2012
 - Visiting instructor at University of Dschang in Cameroon.

Language Technology Consultant (2008*-Present)

- Serving primarily (and currently) in Yaoundé, Cameroon
- Language Technology training and support
- Linguistics, Translation, Literacy, Scripture Engagement, and Publishing
- I serve as a bridge between the language workers and the tech nerds.



Creative Commons Image Source: https://commons.wikimedia.org/wiki/File:Cameroon_map_Lambert-AEA_topographic_with_regions-blank.svg

Overview

- Identification of language endangerment as a pressing issue
- The rise of documentary linguistics as a response of the linguistics community
- Defining language documentation
- Perspectives on why it is important
- Application of LancDoc principles.

Thinking about our audience

- As staff of SIL (mostly), you probably already have a heart for minority linguistic communities.



EGIDS Scale

You know a language begins the slide towards language death once children are no longer learning and using it.

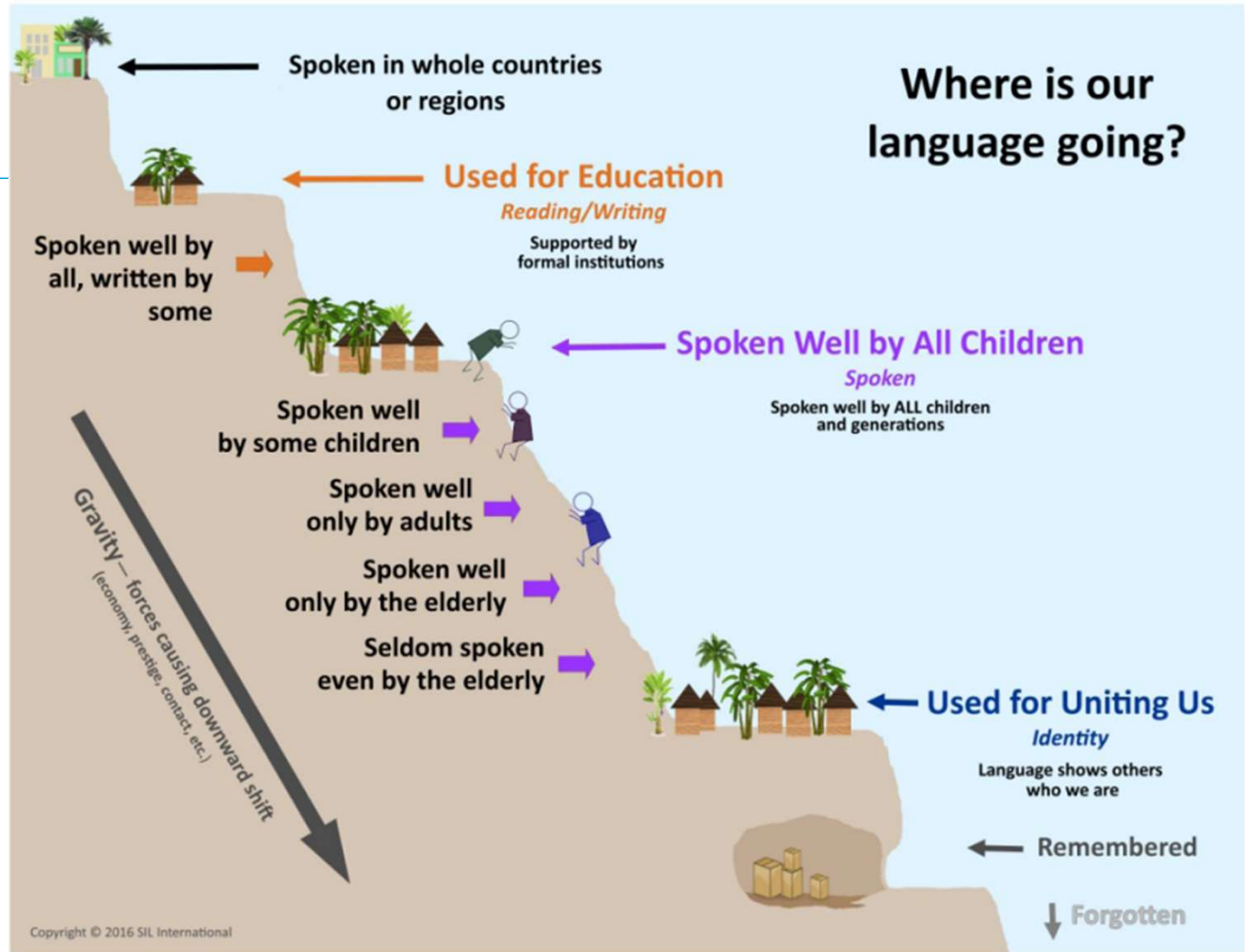


Image source: p25 of Hanawalt, Charlie & Varenkamp, Bryan & Lahn, Carletta & Eberhard, David. (2015). Guide for Planning the Future of Our Language. 10.13140/RG.2.1.4219.0164. https://www.sil.org/sites/default/files/guide-1st_ed_nov_17_2016.pdf

1 000 000+ articles

Polski	Deutsch	Español	Italiano	Nederlands	Português	Sinugboanong	Svenska	Tiếng Việt	中文
العربية	English	Français	مصرى	日本語	Binisaya	Українська	Winaray	Русский	

100 000+ articles

Afrikaans	বাংলা	Eesti	한국어	Latina	Bahaso	O'zbekcha /	Slovenčina	தமிழ்	اردو
Asturiano	Беларуская	Ελληνικά	हिन्दी	Latviešu	Minangkabau	Ўзбекча	Slovenščina	Татарча / Tatarça	粵語
Azərbaycanca	Català	Esperanto	Hrvatski	Lietuvių	Norsk (bokmål ·	Қазақша / Qazaqşa	Српски / Srpski	ภาษาไทย	Հայերեն
Български	Čeština	Euskara	Bahasa Indonesia	Magyar	nynorsk)	قازاقشا /	Srpskohrvatski /	Тоҷикӣ	မြန်မာဘာသာ
Bân-lâm-gú / Hó-ló-oē	Cymraeg	فارسی	עברית	Македонски	Нохчийн	Română	Српскохрватски	تۆرکجه	
	Dansk	Galego	ქართული	Bahasa Melayu		Simple English	Suomi	Türkçe	

10 000+ articles

Bahsa Acèh	Беларуская	Emigliàn–	Hak-kâ-ngî / 客家	Ирон	Lëtzebuergesch	მარტალურნი	Occitan	Qırımtatarca	سنڌي	Volapük
Alemannisch	(Тарашкевіца)	Rumagnòl	語	Íslenska	Ligure	مازرونی	Олык марий	Runa Simi	Ślůnski	Walon
አማርኛ	Bikol Central	Fiji Hindi	Hausa	Jawa	Limburgs	Ming-dĕng-ngŭ /	ଓଡ଼ିଆ	संस्कृतम्	Basa Sunda	吳語
Aragonés	बिष्णुप्रिया मणिपुरी	Føroyiskt	Hornjoserbsce	ಕನ್ನಡ	Lombard	閩東語	অসমীয়া	Саха Тыла	Kiswahili	שייטל
Արևմտահայերէս	Boarisch	Frysk	Ido	Kreyòl Ayisyen	मैथिली	Монгол	ਪੰਜਾਬੀ (ਗੁਰਮੁਖੀ)	Scots	Tagalog	Yorùbá
Basa Bali	Bosanski	Gaeilge	Igbo	Kurdî / كوردی	Malagasy	Napulitano	(پنجابی شاہ مکھی)	ChiShona	தமிழ்த்	Zazaki
Bahasa Banjar	Brezhoneg	Gàidhlig	Ilokano	کوردیی ناوەندی	മലയാളം	नेपाल भाषा	پښتو	Shqip	తెలుగు	Žemaitėška
Basa Banyumasan	ЧӀавашла	ગુજરાતી	Interlingua	Кыргызча	文言	नेपाली	Piemontèis	Sicilianu	සිංහල / Basa Ugi	isiZulu
Башкортса	Diné Bizaad		Interlingue	Кырык мары	मराठी	Nordfriisk	Plattdüütsch	සිංහල	Vèneto	

1 000+ articles

Dzhudezmo /	Aymar	Davvisámegiella	گیلکی	Коми	Lingála	Náhuatlahtōlli	Pfälzisch	Seediq	Tok Pisin	Wolof
די ייִדיש	भोजपुरी	Deitsch	韓国 / 韓語	Перем коми	lojban	Dorerin Naoero	Picard	Seeltersk	faka Tonga	isiXhosa
Адыгэбзэ	Bislama	ትግርኛ	Gungbe	Kongo	Luganda	Nedersaksisch	Къарачай–	Sesotho sa Leboa	СӰУ	Zeeuws
Ænglisc	Буряад	Dolnoserbski	Хальмг	कोंकणी / Konknni	Malti	Nouormand /	малкъар	Setswana	chiTumbuka	Reo tahiti
Anarâškielâ	Чавасано де	Эрзянь	ʻŌlelo Hawaiʻi	کٲشُر	Māori	Normaund	Qaraqalpaqsha	Словѣньскъ /	Türkmençe	
аңсһаа	Zamboanga	Estremeñu	Ikinyarwanda	Kriyòl Gwiyannen	Twi	Novial	Ripoarisch	ᱫᱷᱟᱱᱵᱟᱫᱽ	Тыва дыл	
Armãneasche	Corsu	Furlan	Kabyɛ	କାଶ୍ମୀରୀ	Mirandés	Afaan Oromoo	Rumantsch	Soomaaliga	Удмурт	
Arpitan	Vahcuengh / 話僮	Gaelg	Karampangan	Лакку	Мокшень	පර්දිදි:ဘာဒ်သာ	Русиньскый	Sranantongo	ئۇيغۇرچە	
ܐܘܪܘܩܝܬܐ	Dagbanli	Gagauz	Kaszëbsczi	Latgaju	Latgaj	ဘာသာ မန်	Gagana Sāmoa	Taqbaylit	Vepsän	
Avañe'ẽ	الدارجة	Gikũyũ	Kernewek	Лезги	Lezgi	Li Niha	ᱯᱟᱨᱚᱰᱟ	Tarandine	Vöro	
Авар			සාහසිලිය				Papiamentu	Sardu	Tetun	West-Vlams

Why does language death matter?

- The **scientific** significance
 - Some linguistic phenomena are known only through these languages.
 - Claims about linguistic universals need to be tested against all languages.
 - Clues about the prehistory of peoples are lost.

Why does language death matter?

- The **ecological** significance
 - Diversity is good for the planet.
 - Just as the environment needs biological diversity in order to thrive and to rebound from disaster,
 - So humankind needs diversity of knowledge and ways of thinking in order to thrive on our planet and to adapt to changing situations.
 - The loss of a language diminishes this capacity.
 - Each language encodes unique and valuable knowledge
 - When a language is lost, humankind loses with it a vast store of knowledge about the environment in a particular region of the planet—knowledge that took millennia to develop and that is never likely to redevelop.

Why does language death matter?

- The **legal** significance
 - One tool in combating the loss of languages has been to establish legal rights for language use and development. For example,
 - *United Nations Declaration on the Rights of Indigenous Peoples (2007)*
 - [History](#) on Wikipedia; [Full text](#) on UN site
 - Article 13
 - 1. Indigenous peoples have the right to revitalize, use, develop and transmit to future generations their histories, languages, oral traditions, philosophies, writing systems and literatures.

Why does language death matter?

- The **social** significance
 - Language is an expression of identity; loss of identity is a serious problem for groups and individuals alike.
 - Victims of language loss have been shown to have worse health (see *Healing through language*)
 - “Tribal children given the opportunity to learn their language are happier, healthier human beings. It doesn’t mean their lives are easier. It does mean that their identities are stronger and they are better prepared to face the challenges of being an indigenous person in the modern world.”
— Jacob Manatowa-Bailey

Why Does Language Death Matter?

- The **theological** significance
 - God intends that every people and language should be part of his kingdom (Dan. 7, Rev. 7)
 - Even when a language is going to be lost, there can be spiritual motivations for documenting it:
 - Helping people sustain their identity as a distinct people.
 - Preserving a record of a human language is capturing a facet of God's creative glory.
 - Honoring God by publishing descriptions that put his workmanship on display to a watching world.
 - Honoring God by honoring the people he created in their uniqueness.

The Birth of Language Documentation

- Definitive source on documentation vs description:
 - Nikolaus **Himmelman**, 1998. "Documentary and descriptive linguistics." *Linguistics* 36:165–191.
- Definitions
 - Documentation is “the activity concerned with collecting, transcribing, translating, and commenting on **primary data**” (190) [+archiving]
 - Aim is “to provide a comprehensive record of the linguistic practices characteristic of a given speech community.” Contrasts with description which aims at “the record of a language ... as a system of abstract elements, constructions, and rules.” (166)

Documentation vs. Description

	<i>Documentation</i>	<i>Description</i>
What?	Primary data	Secondary data
How?	Observe, Record, Transcribe, Translate	Analyze, Generalize
Who?	Recording specialists, Literate speakers	Professional linguists
Where?	On site	On or off site
When?	Short term	Long term

How does this compare to traditional descriptive practice?

- Traditional descriptive linguistics is to:
 - Publish a grammar (with phonology), dictionary, and interlinearized texts—the Boasian trilogy.
 - Requires extensive training and years of effort.
 - In practice, few projects produce all the above.
 - What's worse, there has not been a practice of archiving recordings, so that *the primary source material* is not available for:
 - Language revitalization or identity maintenance
 - Supporting descriptive claims that are published
 - Analysis by others to fill descriptive gaps



Language Documentation as a Specialization

Language Documentation as a Specialization

- Language Documentation is the collection, annotation, and archiving of a corpus of raw primary linguistic data.
 - Language Documentation can be a project/goal unto itself, carefully collecting and annotating a broad corpus of language and culture data that will be preserved for later generations.
 - Such an organized corpus could easily become the foundation of linguistic work, language development, or language revitalization (but that is an analytical exercise left to the reader after documentation has been done).

Basic Documentation Tasks

- Make ***archival quality*** digital recordings of the language in a variety of uses (including an extended wordlist).
- Provide metadata documenting the full situation surrounding each recording.
- Transcribe each recording with time alignment.
- Translate each transcription at sentence level and even word level.
- Obtain and document informed consent to share.
- Archive all of the above as a corpus.
- Allow public access as a citable publication.

The three basic tasks

“Language Documentation is concerned with compiling, commenting on, and archiving language documents.” — Himmelmann 1998

1. **Compile (collect)** a sample of recordings of a full range of speech event types
2. **Comment on (annotate)** those recordings
 - E.g., transcription, translation, discussion, situational context, informed consent to share
3. **Archive** the complete corpus of recordings and commentary with an institution that will provide long-term access

“Classic” Language Documentation

Stemming from Himmelmann and Woodbury, “classic” Language Documentation includes:

- recording (audio, video, and photos)
- adding value (annotation and metadata)
- and archiving all of the above.

It stops there. Further analysis and facilitating use of the content are not in part of the expectations.

Austin's Five Activities of Language Documentation

According to [Austin \(2014, 60-61; 2006; 2010, 19\)](#), the ideal language documentation process contains five activities:

- Recording
- Transfer [for mgmt.]
- Adding value
- Archiving
- Mobilization



Language Documentation Methods

(As an Add-on to Your Work)

In the course of your language work, you may:

- Collect texts and linguistic paradigms.
- Compile a wordlist or dictionary.
- Collect recordings and photos.
- Analyze and annotate the above content.

and finally...

- Write papers and give presentations on your findings.

A Moment for Humility

- While the analytical papers you write may make you rich and famous (in the linguistic world), we need to realize that the culturally-relevant content you collected along the way may sometimes be more valuable to the community.
 - ...as a cultural archive
 - ...as a historical record
 - ...as the building blocks for language revitalization
 - ...as a standardizing force
 - ...as a source for more advanced work

Language Documentation as an “Add-on”

- Practically, Language Documentation methodologies on a small scale promote proper stewardship of the primary data collected in the course of research.
- Even elicitations that were narrowly selected to respond to a specific research question could still help to serve future cultural and language-development purposes.
- Every language worker could benefit from a “minor” in Language Documentation techniques.

Primary Recordings

- A blind spot in our practice of field work to date involves the primary recordings:
 - Virtually every project produces them, but
 - Virtually no project has published or preserved them
- That is no longer acceptable practice:
 - Language communities are calling for them
 - Academic community is calling for them
- We should want to comply because language documentation has so many ***potential benefits***, which we look at next.

Including Data Enhances Your Credibility:

- Which sounds better?
 - “Nchane exhibits a contrasting /e/ and /ɛ/ such as in the words *ge* ‘do’ and *ge* ‘not’.”
 - “Nchane exhibits a contrasting /e/ and /ɛ/ such as in the words *ge* ‘do’ and *ge* ‘not’ as heard in recordings W-012 (at 0:30) and W-032 (at 0:21) in the linked corpus.”
- Including referenced source data in a corpus enriches the reading, allowing the reader to verify your work. Including ALL source data gives the reader the chance to test your conclusions AND do further research.

Reproducible Linguistics

- Science is about testing and retesting results.
- A chemistry experiment may be reproducible with chemicals found in the laboratory, but minority linguistics experiments can not be easily repeated without access to the raw data.
- This push for publishing both the data and analysis is called “reproducible research” and has been extended to linguistics.

Berez-Kroeker, Andrea L et al. (2018). “Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field”. In: *Linguistics* 56.1, pp. 1–18. doi: 10.1515/ling-2017-0032.

Reproducible Linguistics

- If a linguist makes a claim that a skeptic can only reproduce by spending three decades working in the same language community in the same sociolinguistic and fieldwork conditions, the original claim is really only reproducible in principle. [...] Our view is that it is not healthy for linguistic descriptions to be supported by examples that cannot be reproduced except by doing one's own fieldwork [...] it may be *research* and it may be *important*, but unless enough details of the utterances in context are made available so that it can be subjected to true reproducibility tests by skeptics, it isn't Science.

(Gezelter paraphrased by Berez-Kroeker et al (2018:4))

-
- The contribution of Simons (2008), Boerger (2011), and Boerger et al. (2019) represents two major elements: (1) a recommendation for the minimal contents of a language documentation corpus and (2) canonization of an oral-first workflow of annotation.



The Language Documentation Process

1. Recording — of media and text (including metadata) in context

- *Elicited Lists*
 - Wordlists and Grammatical Paradigms
- *Communicative Events*
 - Interactive Discourse, Narrative, Procedural Discourse, Oratory, Description, Drama, Poetry, Songs, Formulaic Discourse, Language Play, Unintelligible Speech
- *Analytical Discussions* about any other (meta)content.
- Written Works by Local Authors
- Events, Activities, Images, Locations, Maps, Artifacts, etc.
- *Descriptive Metadata* about all content collected.



Photo by [zhenzhong liu](#) on [Unsplash](#)

Other sampling dimensions

- The choice of speakers should involve sampling as well. Universally applicable:
 - Gender
 - Age
- Relevant in some situations:
 - Social stratum
 - Education level
- Also sample regional varieties

2. Transfer — to a data management environment

SayMore by
SIL International


- Gathering
- Converting
- Organizing
- Annotating
- Checking Coverage
- Archiving

The screenshot displays the 'French Transcription - SayMore 3.2.17 (Release)' application window. The interface is divided into several sections:

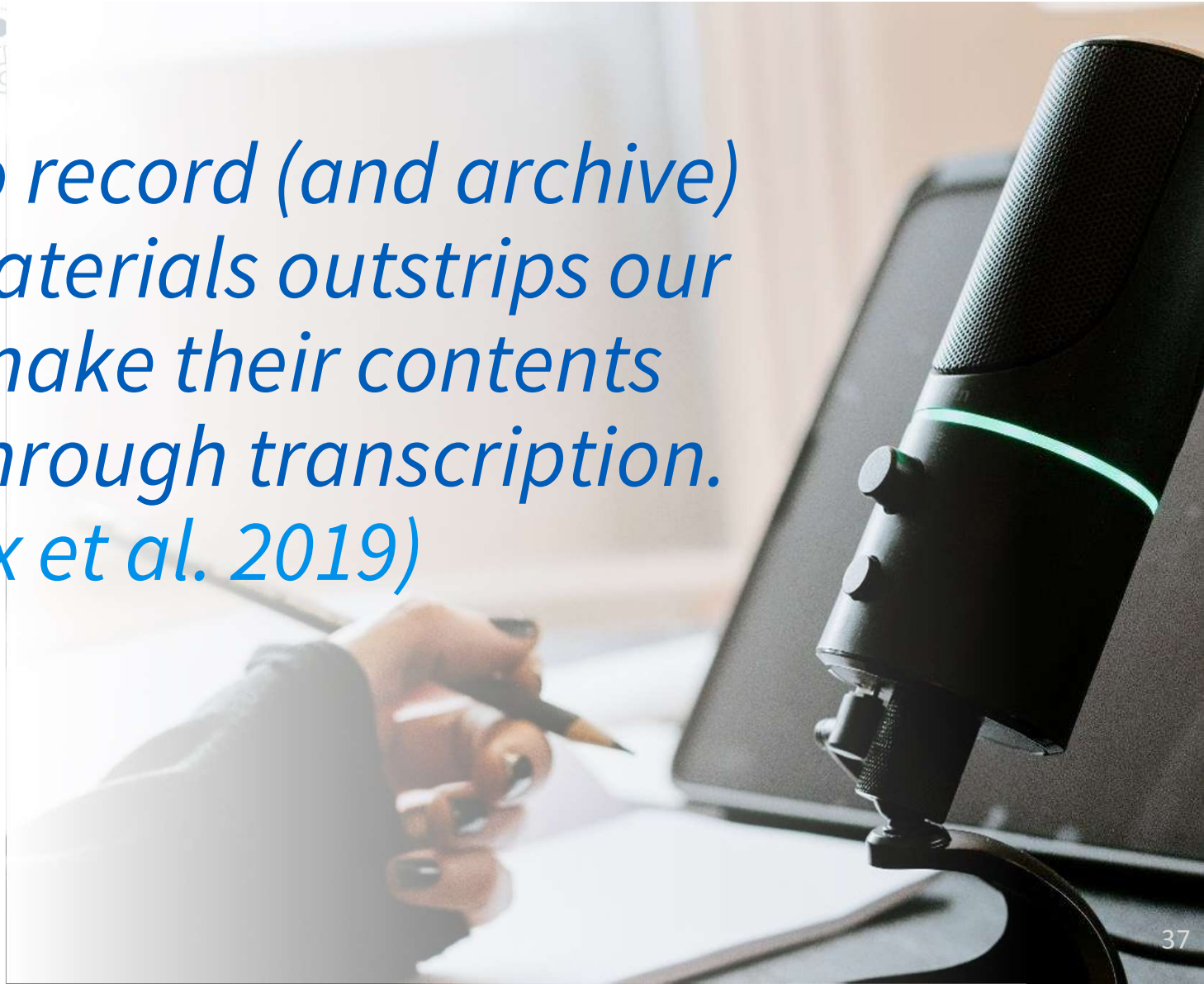
- Menu Bar:** 'Projet', 'Session', 'Personne', 'Aide'.
- Navigation Bar:** 'Projet', 'Sessions', 'Personnes'.
- Sessions List (Left):** A table with columns 'ID', 'Titre', 'Étapes', and 'État'. The selected session is 'Pourquoi un mètre m...'. Other sessions include 'Intro2', 'Introduction', 'Message for Camero...', 'Michel', and 'Orthography Sadembuo'.
- File List (Right):** A table with columns 'Nom', 'Type', and 'Date de modification'. It lists various files such as 'Pourquoi un mètre mesure 1m.session', 'Pourquoi un mètre mesure 1m_Source_01.mp4', and several audio and annotation files.
- Session Detail View (Bottom):** A form for editing session details. Fields include:
 - ID:** 'Pourquoi un mètre mesure 1m'
 - Date:** '9/ 1/2019'
 - Titre:** (empty)
 - Personnes:** 'Matthew Lee (Transcriber); Bruce Benamram (Speaker);'
 - Genre:** 'scours formulé'
 - Accès:** '<inconnu>'
 - Situation:** (empty)
 - Description:** (empty)

3. Adding value — the transcription, translation, annotation and notation and linking of metadata to the recordings

- *Written Transcription* in the orthography or IPA
- *Written Translation* into a language of wider communication
- Interlinearization?



*Our ability to record (and archive)
language materials outstrips our
ability to make their contents
accessible through transcription.
(Cox et al. 2019)*





Transcription Bottleneck

- ◎ The extreme workload of written transcription (the *transcription bottleneck* (Seifart et al. 2018, 1)) means that large amounts of language documentation content may never be “ready” for distribution.
- ◎ At best, the unfinished content may be hidden in the archive obscurely.
- ◎ At worst, staff and resources can be reassigned, and unfinished content may be lost to history.

The problem

- Small language communities are losing their languages faster than linguists using conventional methods can document them.
 - Using written methods creates a transcription and analysis bottleneck.
- What should we be doing in response?
- A possible solution has arisen from combining two ideas.
- We'll look at a proposed partial solution to the transcription bottleneck (BOLD) later.



Breaking the Transcription Bottleneck

Developing a BOLD approach

- A team at SIL developed a method called:
 - **B**asic **O**ral **L**anguage **D**ocumentation
- In place of the traditional corpus approach:
 - Compile, Transcribe, Annotate, Archive
- We build an oral documentation corpus:
 - Compile, Comment orally, Archive
- An advantage is that while a linguist may be the catalyst, even non-linguists (e.g. community members) can be mobilized to do the work of compiling and commenting, thereby speeding the work and often increasing accuracy.

BOLD Oral Annotation

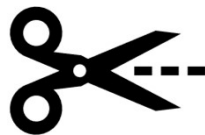
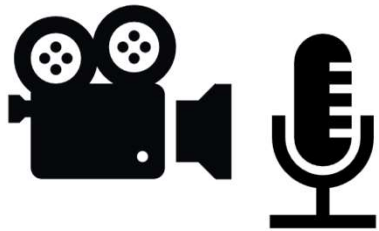


- In Context

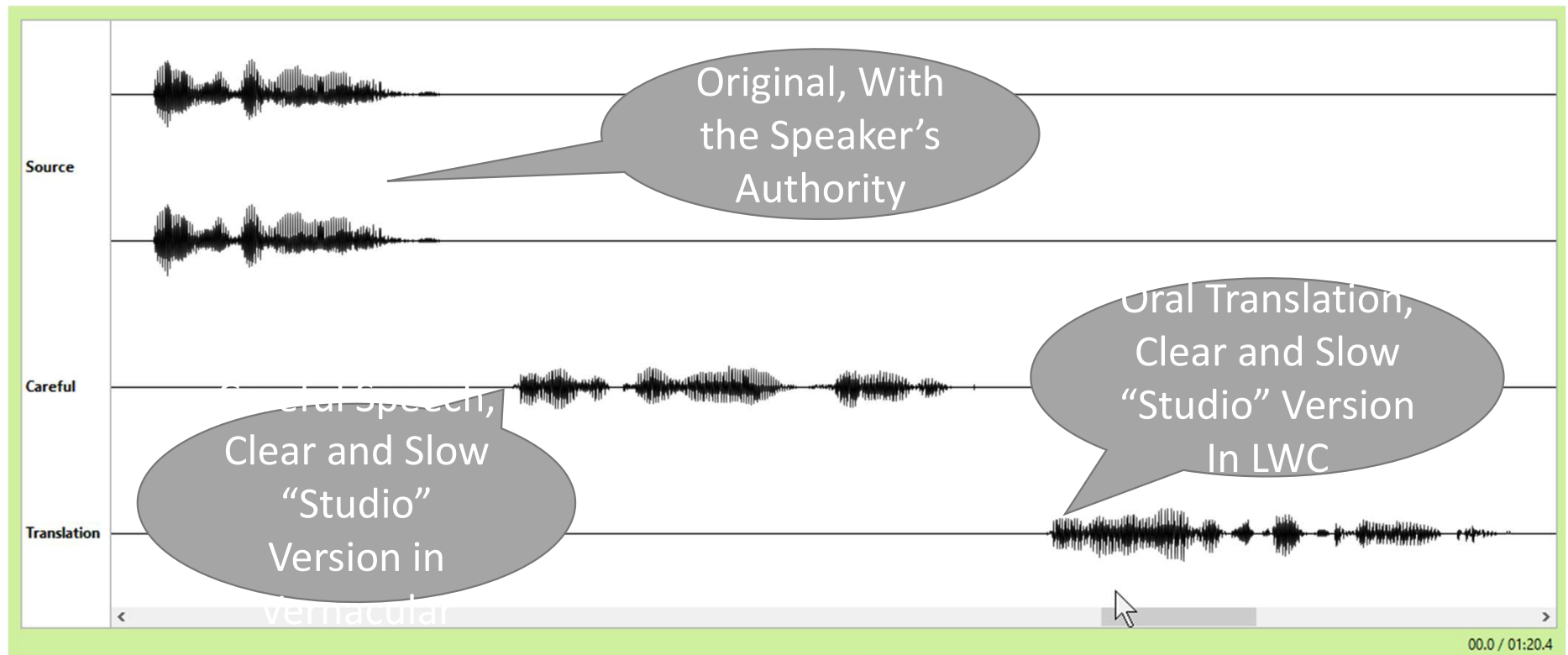
- Automated Cuts at Breaks, with Fine Tuning

- Each Phrase Repeated Slowly by a Qualified Speaker

- Each Phrase Free Translated by a Bilingual Speaker



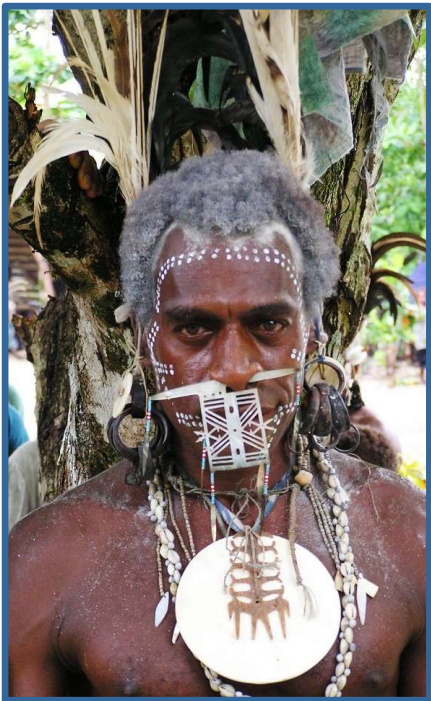
BOLD Annotation Creates Aligned Oral Transcription and Translation



Adding written transcription and translation

- Oral transcription and oral translation are the source for written transcription and translation
 - Either done immediately and added to the documentary corpus
 - Or done later (even as a different project by different people) as the basis for a new descriptive corpus with links back to the sources in the documentary corpus

Further Reading



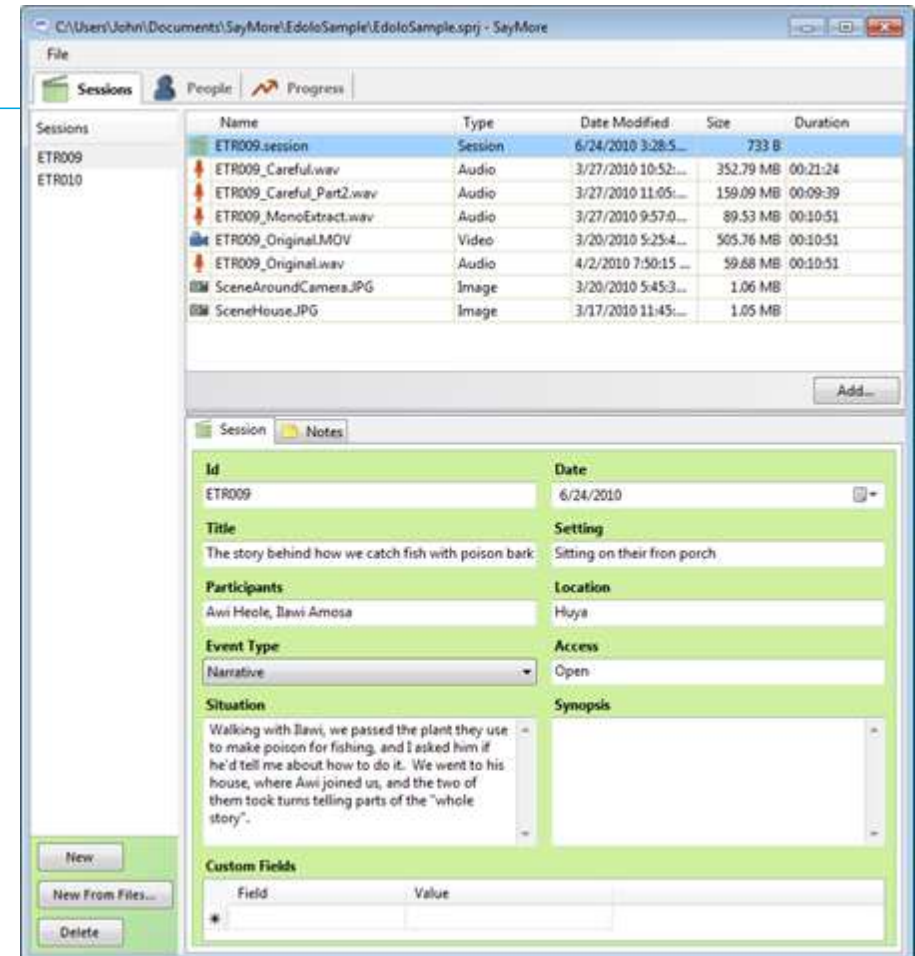
- Woodbury, Anthony. 2003. “Defining documentary linguistics.” In Peter Austin (ed.), *Language Documentation and Description 1*. HRELP, SOAS.
- Bird, Steven and Gary Simons. 2003. “Seven dimensions of portability for language documentation and description.” *Language* 79:557-582.
- Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel (eds.). 2006. *Essentials of Language Documentation*.
- Boerger, Brenda H., Sarah Moeller, Will Reiman and Stephen N. Self. 2016. *Language and Culture Documentation Manual*, Leanpub.
- Reiman, D. Will. 2010. Basic Oral Language Documentation. *Language Documentation and Conservation*, Vol. 4:254-268.
- Boerger, Brenda H. 2011. To BOLDly Go Where No One Has Gone Before. *Language Documentation and Conservation*, Vol. 5:208-233.

Tool for a lang doc corpus

SayMore

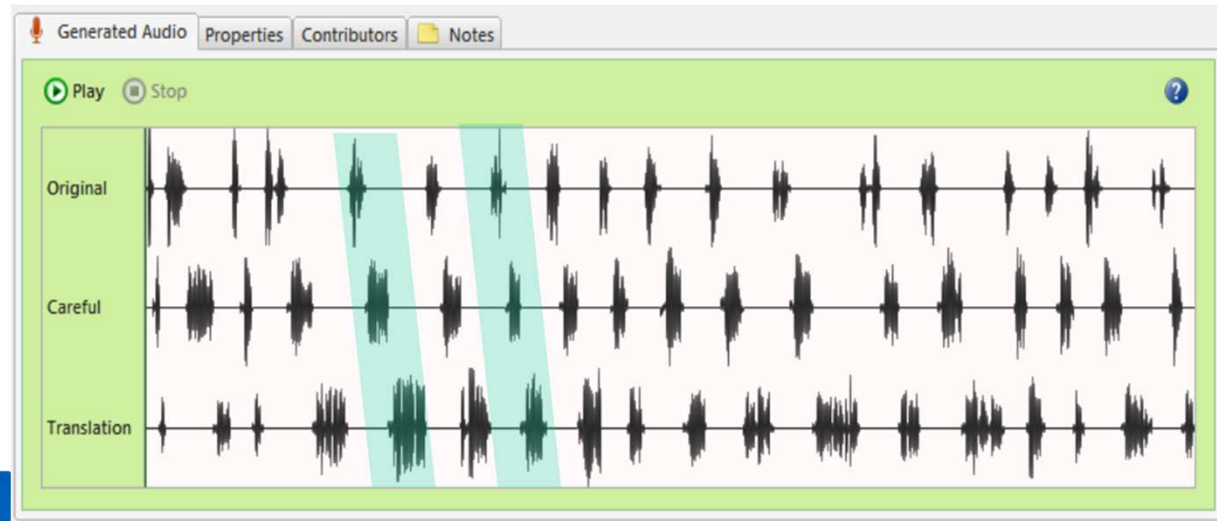
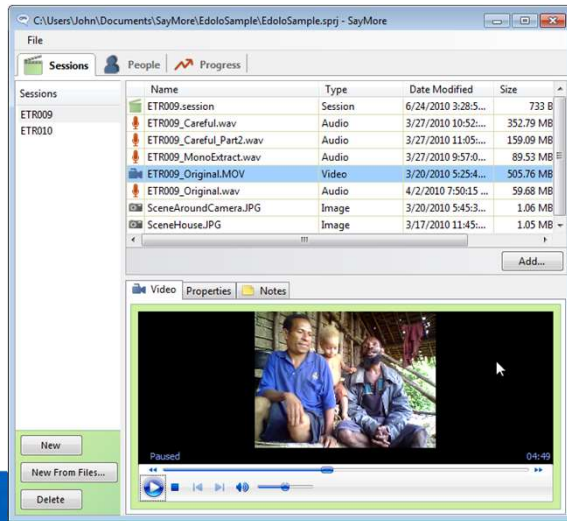
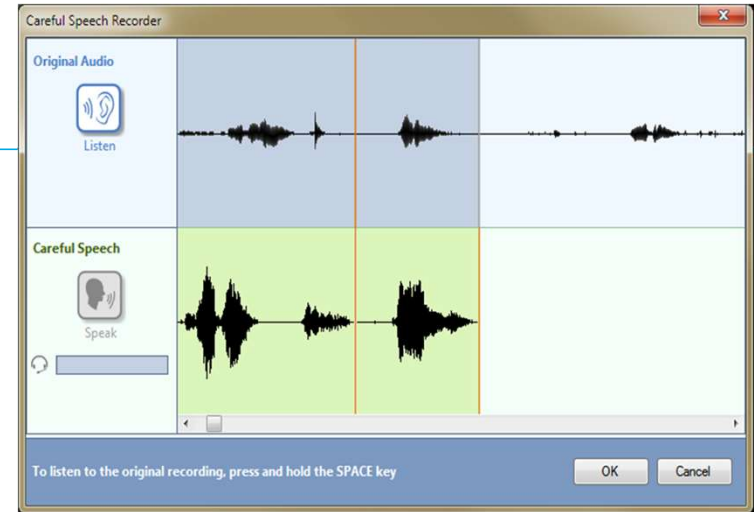
Lang Doc Productivity

- Organize
 - Session data
 - People data
 - Media files
- Track progress
- Do transcription
- Archive results
- Download v. 3.1.5
- <https://software.sil.org/saymore/>
- <https://github.com/onset/laMETA/releases>

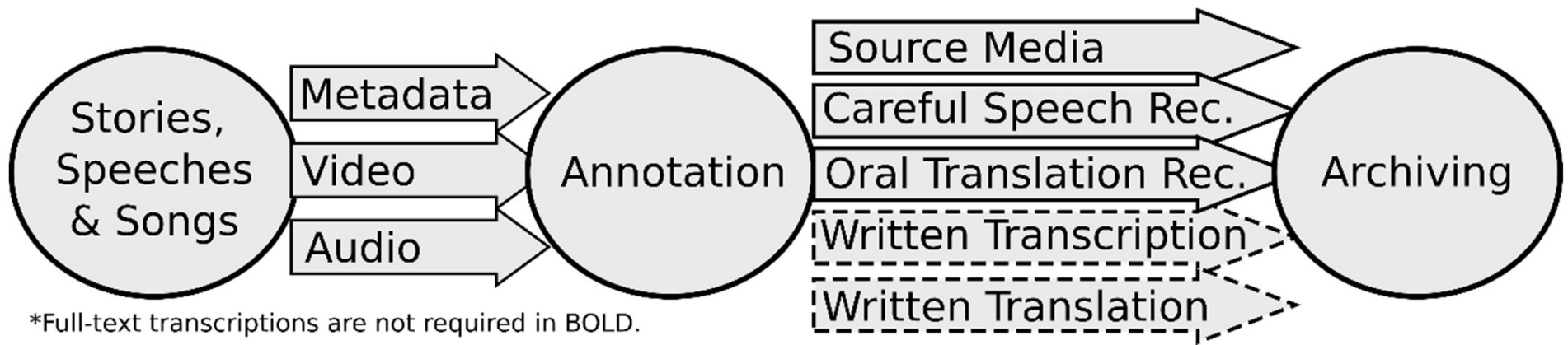


More features

- Auto segments files
- Handles video
- Automates the BOLD respeaking process
- Creates staggered audio for lg. learning
- Moeller (2014)



BOLD Corpus is Ready to Archive



4. Archiving — creating archival objects and assigning them access and usage rights;

SayMore facilitates archival of quality primary recordings, full metadata, and annotation to preserve aspects of the language and cultural heritage for the long term.

5. Mobilization — creation, publication and distribution of outputs in a range of formats for a range of users and uses.

Mobilization in language documentation means working with speaker communities to produce products from language documentation that can be used to counter language endangerment (Nathan 2006, 363).



Archives and Vaults



Submissions to the Seed Vault

- Samples are cleaned, decontaminated and carefully frozen so that they can last hundreds of years.
- Individual items are clearly labeled in bags and boxes.
- Metadata is created with sources, dates, names, and species and added to a central index.
- The boxes are placed in specific locations so that they can be found.
- The natural ice helps keep temperatures cool, but the staff watch for other risks.

Linguistic Archives are similar to the Seed Vault

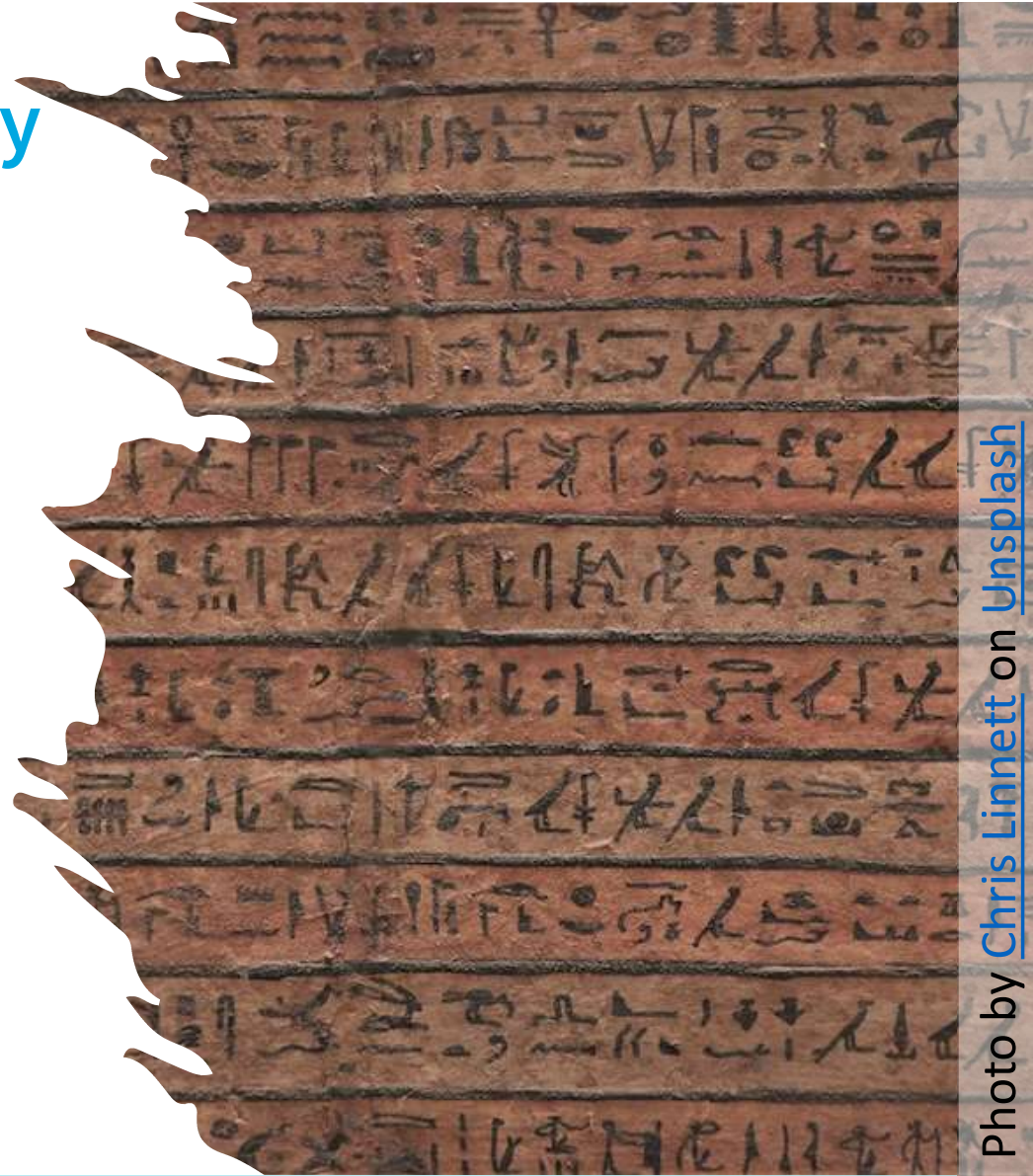
- Content is prepared in the best quality possible.
- Individual items are clearly labeled in files and folders.
- Overview metadata is collected on content, language, situation and date.
- The materials are submitted to relevant sections of the archive and an index is maintained.
- The archivists actively work to protect the storage from fire, water, mechanical failures, and corruption.

A paradox of writing history

The more advanced the writing technology, the more it can store but the less durable the written product.

From most durable to least durable:

- Clay tablets and stone (2,000 yrs+)
- Vellum
- Papyrus
- Paper
- Digital word processing (tens of years?)



Digital Storage media are ephemeral

- Life expectancy of digital storage media:
 - Magnetic tape and disks: 10 to 20 years in low humidity.
 - CD-R (write once)
 - Manufacturers say: 100 to 200 years
 - Independent lab says: 30 years
 - CD-RW (write many times)
 - Manufacturers say: 25 years
 - Traditional “Spinny” Hard Drives
 - 3-5 years depending on use.
 - SSD drives
 - It may be too early to know.

Forms contrasted by function

- **Working form**
 - The form in which information is stored as it is created and edited.
- **Presentation form**
 - The form in which information is presented to the public.
- **Archival form**
 - The form in which information is stored for access long into the future.

The Three Forms

Adapted from <https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html>

	Archival Form	Working Form	Presentation Form*
Photos	.TIFF	Photoshop File	.JPG, .PNG
Audio	.WAV, .AIFF	Audacity Project	.MP3, .ACC
Video	.AVI, .MXF, .MOV	Final Cut Project	.MP4, WEBM
Documents	Unicode Text, OpenDocument Text (ODF), PDF/A, Rich Text Format	Word DOCX, Publisher, PowerPoint, Pages, InDesign, Google Docs, web pages	

SIL's digital archive

- REAP — Repository for Electronic Archiving and Publishing
- SIL's digital archiving system:
 - Released in 2011 as link on *Insite* home page
 - Based on DSpace, an open-source system used by hundreds of universities
 - Uses a web browser to supply metadata through forms and upload new resources
 - Includes RAMP — a tool for offline acquisition
 - Deposits marked as “Public” are automatically shared via sil.org and OLAC

July 21, 2021

Dear Svaldbard Seed Vault,

I am in desperate need of some more seeds to plant this year. Can you please send some seeds to help save my farm? I've enclosed a list of the seeds I need.

Thanks in advance,

D. A. Fon, Bafia, Cameroon

Where the analogy to seeds breaks down.

- If someone removes the some of the seeds, the vault now holds less seeds.
- One of the advantages of digital media is that it can be copied without the vault losing the original, so why would the archive still need to limit access?

However...

- Every time the data is accessed, the drives edge closer to failing.
- Drives must be tested for data integrity, which adds to wear.
- Access increases the risk of corruption, viruses, and hacking.



Discoverability



“But look, you found the notice, didn’t you?”
“Yes,” said Arthur, “yes I did. It was on display in the bottom of a locked filing cabinet stuck in a disused lavatory with a sign on the door saying ‘Beware of the Leopard.’”

— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*



Open Language Archives Community

www.language-archives.org

- OLAC is an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by:
 - Developing consensus on best current practice for the digital archiving of language resources
 - Developing a network of interoperating repositories and services for housing and accessing such resources
- Founded in December 2000
 - Now has 61 participating archives
 - ~400,000 language resources in catalog

A focused search strategy

<http://search.language-archives.org>

The screenshot shows the OLAC Language Resource Catalog website. The browser address bar displays `search.language-archives.org/index.html`. The page title is "Prototype: OLAC Language Resource Catalog".

Search for language resources [input field] [go]

IMPORTANT: This is just a PROTOTYPE!

- What is a DLA PROTOTYPE?
- Navigating the Catalog**
 - Catalog Home
 - Search Strategies
 - Advanced Search
 - New: Records recently added or modified
- Quick Links**
 - Browse by Language
 - Browse by Country
 - Browse by Linguistic Field
 - Browse by Linguistic Type
 - Browse by Language Family
- Contacts**
 - Email Us
- More information**
 - OLAC Homepage
 - OLAC FAQ
 - Participating Archives

Powered by the DLA

This catalog, developed by the **Open Language Archives Community (OLAC)**, provides access to a wealth of information about thousands of languages, including details of text collections, audio recordings, dictionaries, and software, sourced from dozens of digital and traditional archives.

Browse the OLAC records by Geographic region or by Language:

- English (3509)
- Spanish (2748)
- Yuracare (1269)
- Beaver (1044)
- French (1003)
- Ixóó (845)
- Bora (748)
- Ocaina (678)
- Chhintange (673)
- Kriol (578)
- Portuguese (575)
- Nepali (559)
- Motlav (493)
- Trumai (470)
- Achinese (435)
- Saiba (429)
- Occitan (post 1500) (426)
- Afrikaans (408)
- South West Bay (405)
- Bahinemo (401)

View more...

Browse the OLAC records by Archive:

- A Digital Archive of Research Papers in Computational Linguistics (3280)
- ATILF Resources (9)
- Aboriginal Studies Electronic Data Archive (ASEDA) (707)
- Academia Sinica Collections (3)
- African Language Materials Archive (53)
- Alaska Native Language Center Archive (24)
- Archive of the Indigenous Languages of Latin America (100)
- Audio Archive of Linguistic Fieldwork (199)
- Boiste (1)
- CHILDES Data repository (275)
- CRDO archive (161)
- Central Institute of Indian Languages: Publications (345)
- Centre de Ressources pour la Description de l'Oral (CRDO) (3752)
- Magoria Books' Carib and Romani Archive (2)
- Multimodal Learning and teaching Corpora Exchange (25)
- ODIN - The Online Database of Interlinear Text (710)
- Oxford Text Archive (1264)
- Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (6755)
- Perseus Digital Library (1451)
- SIL Language and Culture Archives (22377)
- Surrey Morphology Group Databases (2)
- Survey of California and Other Indian Languages (2448)
- TALKBANK Data repository (182)
- The LDC Corpus Catalog (486)
- The LINGUIST List Language Resources (2039)
- The Natural Language Software Registry (69)
- The Rosetta Project: A Long Now Foundation Library of Human Languages (1553)

Sort by:

- Possible Sorts: all
- Title [a-z][z-a]
- Id [a-z][z-a]
- Date [a-z][z-a]

Browse by:

- Archive** browse
 - SIL Language and Culture Archives 22377
 - IMDI to OAI bridge 14883
 - Graduate Institute of Applied Linguistics Library 8176
 - Ethnologue: Languages of the World 7413view more...
- Online** browse
 - Yes 49386
 - No 44554
- Subject language** browse
 - English 3509
 - Spanish 2748
 - Yuracare 1269
 - Beaver 1044view more...
- Language family** browse
 - Austronesian 11969
 - Malayo-Polynesian 10730
 - South American Indian 8926
 - Indo-European 8299view more...
- Geographic region** browse
 - Americas 21252
 - Asia 16047
 - Pacific 14954

A Google search strategy



Web [+ Show options...](#) Results 1 - 10 of about 12,000 for bar

[Resources in and about the Barai language](#) - 11:02am

Barai grammar highlights. Olson, Michael Leon. 1975. Canberra, Australia: Australian National University. oai:gial.edu:28481; **Barai** sentence structure and ...

www.language-archives.org/language/bbb - [Cached](#) - [Similar](#)

Use any ISO 639-3 code at end of URL

[OLAC Record: oai:paradisec.org.au:TD1-P034](#)

Barai Grammar con't from tape P29 -- 3. More **Barai** vocabulary. -- Side 2 -- 4. Orokaiva (Hamara Village) Lexico Stats. -- 5. Kwena (Managalasi) -- 6. ...

www.language-archives.org/.../oai:paradisec.org.au:TD1-P034 - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [🗕](#)

[+ Show more results from www.language-archives.org](#)

[SIL Bibliography: Barai grammar highlights](#)

"**Barai grammar** highlights." In T. E. Dutton (ed.), Studies in languages of central and south-east Papua, 471-512. Pacific Linguistics C, 29. ...

www.ethnologue.com/show_work.asp?id=11852 - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [🗕](#)

by ML Olson - 1975 - [Cited by 6](#) - [Related articles](#) - [All 2 versions](#)

[Ethnologue report for language code: bbb](#)

Austin's Five Activities of Language Documentation

According to [Austin \(2014, 60-61; 2006; 2010, 19\)](#), the ideal language documentation process contains five activities:

- Recording
- Transfer
- Adding value
- Archiving
- Mobilization

Austin's Five Activities of Language Documentation

To use Nathan's (2006, 363) words:

- © “This ~~chapter~~^{presentation} assumes that you hope that some of your fieldwork results will one day be applied to the maintenance, strengthening, or revitalization of the visited community's language.”

Revitalization

- Language Documentation projects must store archive-quality content.
- Archival-quality materials (AVI, TIFF, and RTF) are NOT the compressed convenient formats used on modern devices.
- While use of these materials is an exercise for the community, revitalization requires use of those materials in ways familiar to the speakers.
 - Today, that means websites, apps, and media.

Archives and Vaults

- Vaults clearly serve a purpose, protecting valuables from the elements, and ourselves for the long term.
- But...we can't forget:
 - What is necessary to turn around the decline of an endangered language.
 - How can these riches be shared now?

https://www.firstvoices.com/

Stz'uminus Stories

Filter Items

Title


Resource Type

[Reset](#) [Filter](#)



'i tsun tse'wutu
I am making a sweater.

[CONTINUE TO STORY](#)



Brandon's priceless moment

[CONTINUE TO STORY](#)



Georgina's First Win

[CONTINUE TO STORY](#)



Joe 'i' Silu Xut'ukw.
Joe and Silu Carving.

[CONTINUE TO STORY](#)



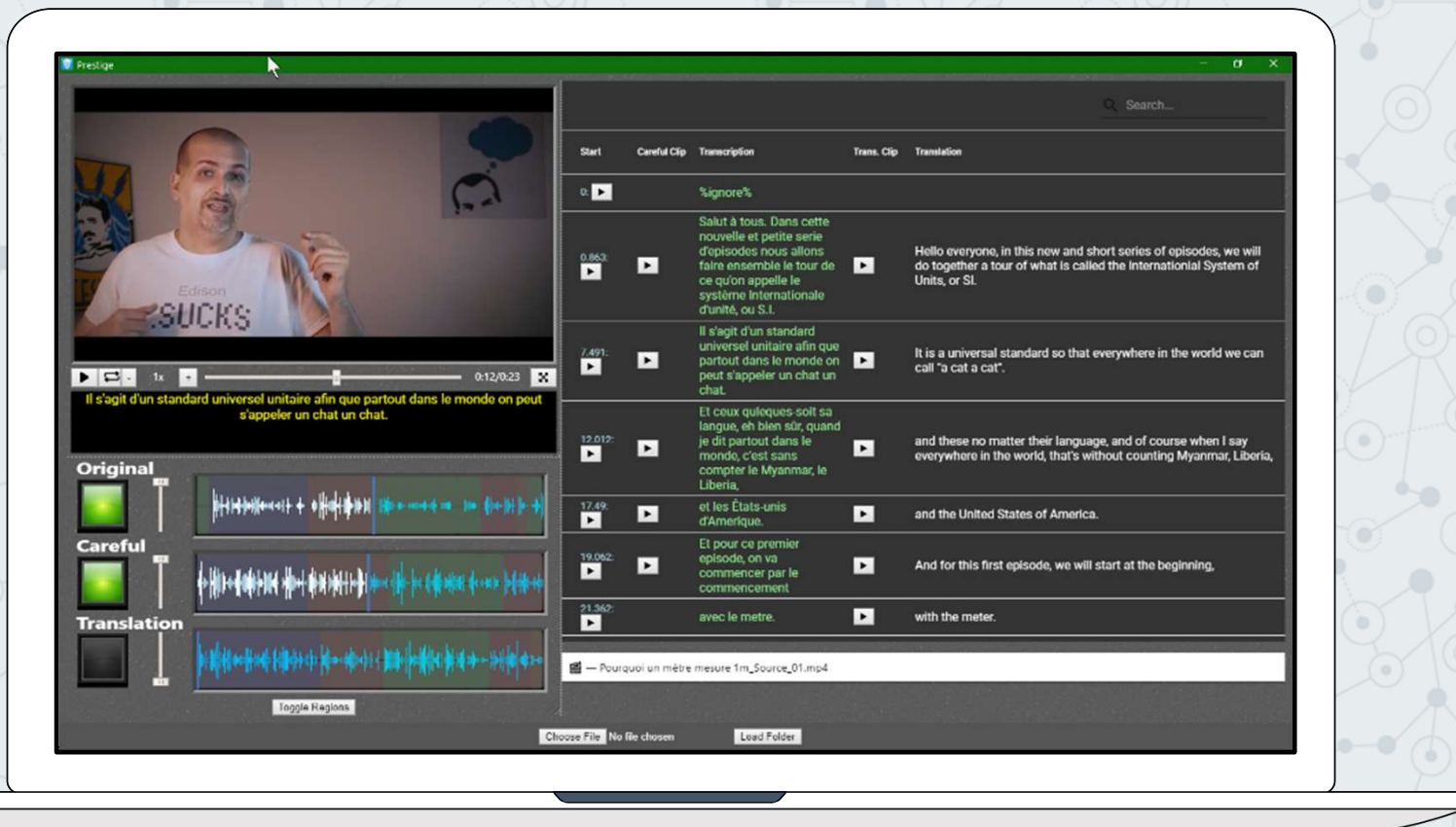
Joe and Silu Canoe Pulling

[CONTINUE TO STORY](#)



Joe and Silu History Lesson.

[CONTINUE TO STORY](#)



Prestige Desktop Prototype

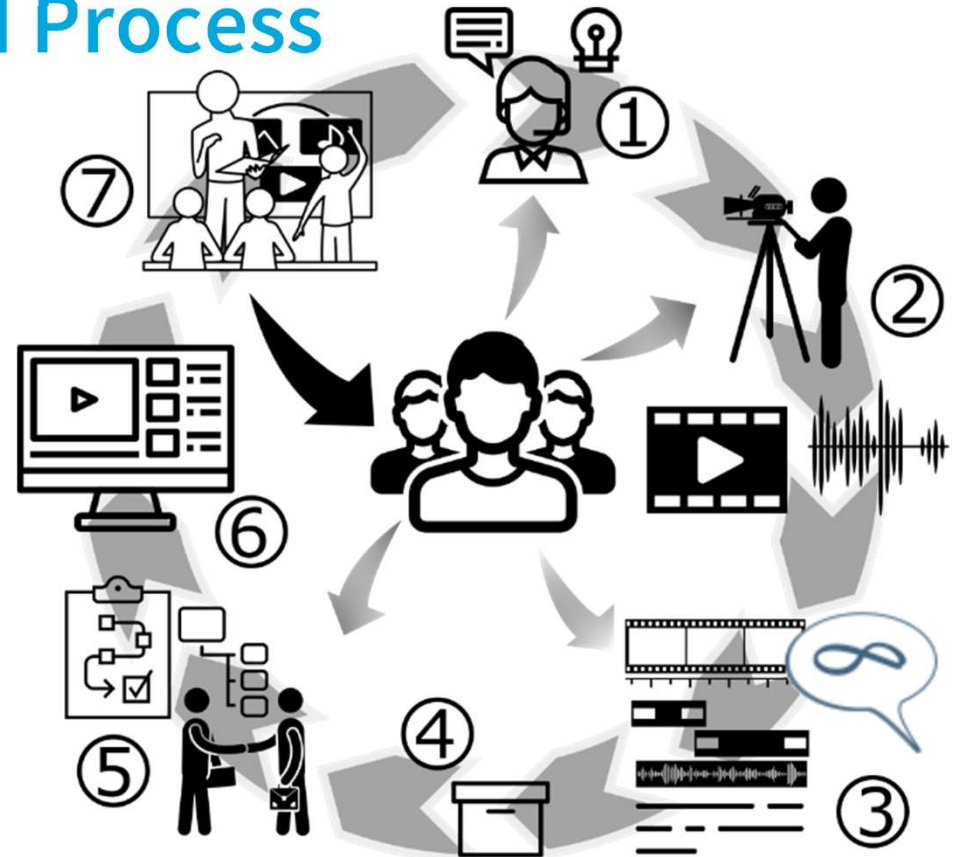
Language Documentation with Prestige A Community-Oriented Process

1. Speakers share stories and songs
2. Performances are recorded in video or audio
3. Recordings are BOLD annotated and translated in SayMore
4. Raw data and annotations are archived for future generations



Language Documentation with Prestige A Community-Oriented Process

5. Community organizes and categorizes content into lessons
6. Content and tagging are imported into Prestige
7. Multimedia content can be used interactively at home or in the classroom for heritage and language learning



A signpost stands in a snowy, mountainous landscape. The signpost has a white rectangular sign with a black border. The sign contains the text 'Language Documentation:' at the top, followed by a blue arrow pointing left and the word 'Resources', and then the word 'Archives' followed by a blue arrow pointing right. The signpost is supported by two black vertical posts. In the background, a long, narrow, dark building is partially buried in snow. The sky is a clear, pale blue.

Language Documentation:

← Resources

Archives →



**How can an annotated
corpus be used?**

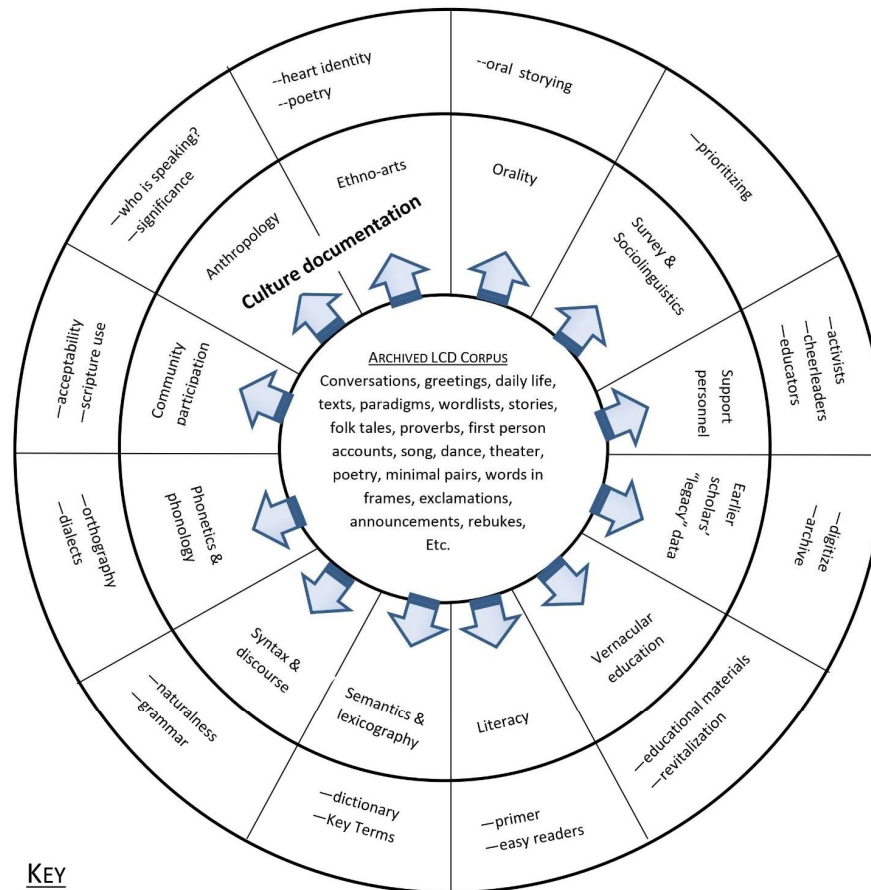
1. Academic uses

- BOLD corpora will serve the *academic community* as a basis for:
 - ***Linguistic description*** — providing primary data for the analysis of phonology, grammar, texts, lexicon (even after the language is gone)
 - ***Linguistic training*** — providing data for examples, problems, and theses
 - ***Transcription and analysis***—easier starting from an orally transcribed corpus

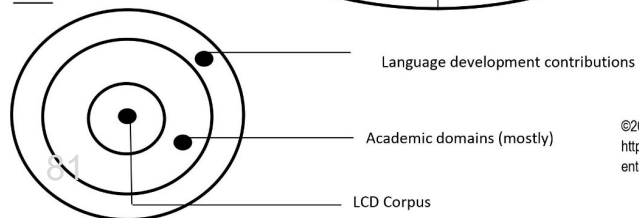
2. Bible translation uses

- A lang doc corpus supports BT because...
 - **Translation personnel**—Are more easily identified while working with community on language
 - **“Comprehensible input”**—From corpus helps in expat’s language learning
 - **Translation naturalness**—Is directly derived from analysis of discourse genres in the corpus
 - **Translation key terms**—These can be revealed in the texts collected as part of the corpus
 - **Language revitalization**—Can be a result of sharing documentation products, such as illustrated story books
 - Contributes to well-being and **strengthens identity**
 - Gives communities time to **respond to the gospel**

Uses for a Language and Culture Documentation Corpus



KEY



©2016 Boerger, Self, Moeller, and Reiman
<https://leanpub.com/languageandculturedocumentationmanual>

3. Church uses

- BOLD corpora will serve the *worldwide church* as they relate to minority language communities to support:
 - ***Language learning*** — providing comprehensible input through oral transcription and translation
 - ***Oral storytelling approaches*** — providing models for oral performance in the language

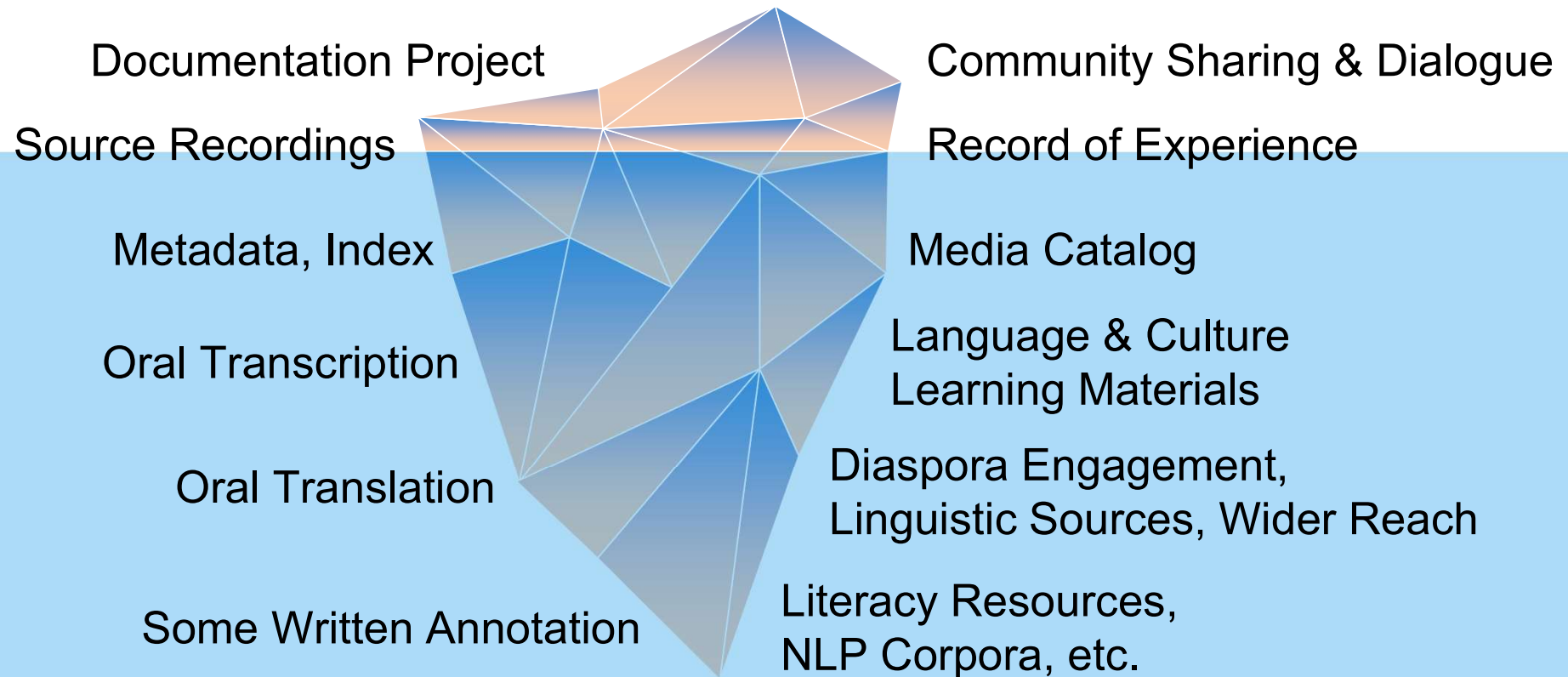
4. Govt & NGO uses

- The services of *governmental and NGOs* will benefit from BOLD corpora to support:
 - ***Language status*** — offering tangible evidence of a people and its language
 - ***Literature & education*** — providing source material for new literature and educational materials for schools and the community
 - COVID-19 materials

5. Language Community uses

- *Minority language communities* will benefit from BOLD corpora since they provide a basis for:
 - ***Heritage preservation*** — saving a record of traditional knowledge and of a group's identity as a people
 - ***Language revitalization*** — providing source material to help people learn their language or learn it better

Language Documentation: Sounding the Depths of a Language Through Annotation



Uncertain times uses

- When **community access is limited**, (like during a pandemic) oral documentation already gathered or quickly gathered can be used for analysis and language learning at a distance
- A team about to leave a project due to sickness or family needs can create a corpus to help **jump start** a new team
- **Uninterested communities** can end up being engaged about their language, language development, and possible translation through interactions during a documentation project, which is not as threatening as BT, perhaps

Conclusion

- A language documentation corpus can be developed in a fairly short period of time by using a purely oral approach.
- Easy access to these corpora will:
 - Address many needs of the scientific community, the worldwide church, governments and NGOs, language communities, and contributes to translation.
- Archiving such corpora will:
 - Ensure long-term access to the recordings, even if the language one day falls silent.