
SOAS guest lecture, 14 October 2020

Language documentation and language description: theory and practice

Peter K. Austin
Department of Linguistics
SOAS, University of London

© Peter K. Austin 2020

Creative commons licence

Attribution-NonCommercial-NoDerivs

CC BY-NC-ND

www.peterkaustin.com

Overview

- A bit of history
 - Some terminology and definitions
 - Relationships
 - The future?
 - Conclusions
-

A bit of history

- In the 18th and 19th centuries the study of language was dominated by historical and comparative considerations (diachrony), especially the reconstruction of past histories of languages and their classification into families
 - Dominance of ‘tree model’ of relationships (cf. ‘wave theory’)
 - Data primarily came from books (especially classical languages, e.g. Sanskrit, Ancient Greek, Gothic Bible)
 - Interest in ‘exotic’ languages with data from missionaries, explorers, travelers, colonial officers (cf. ‘armchair linguists’)
 - Following Frazer, Morgan et al. use of questionnaires and written correspondence with data collectors
-

Man, his relationships, etc.

Aunt	Kgavira, Kaggajitsee
Baby	Aathee, abbala
Blackfellow	Amimouf
Blackwoman	Inialao
Boy	—
Brother	Wosida
Brother-in-law...	...	Wosida , Ammeetha
Child	—
Daughter	Abbala
Daughter-in-law	...	—
Father	Amima
Father-in-law	Amibarnas
Girl	Woorinias
Granddaughter	...	—
Grandfather	Amimee
Grandmother	An'dharree
Grandson	—
Husband	Aridao
Man	Amimouf
Mother	Aagao
Mother-in-law...	...	Kgavira
Nephew	Wagavira
Niece	Wagavira

Daisy Bates,
Western Australia
vocabularies

A bit of history

- Some researchers became interested in local folklore and ‘dialects’, which were seen as disappearing in the face of national (standard languages and cultures), e.g. Grimm brothers
- Beginnings of fieldwork with face-to-face interviews with “best speakers” NORM (non-mobile old rural men) – dialectology. **Method:** long questionnaire to elicit single word answers. **Goal:** creation of linguistic atlas showing geographical distribution of forms
- Began and flourished in Germany, France, Italy in 19th century

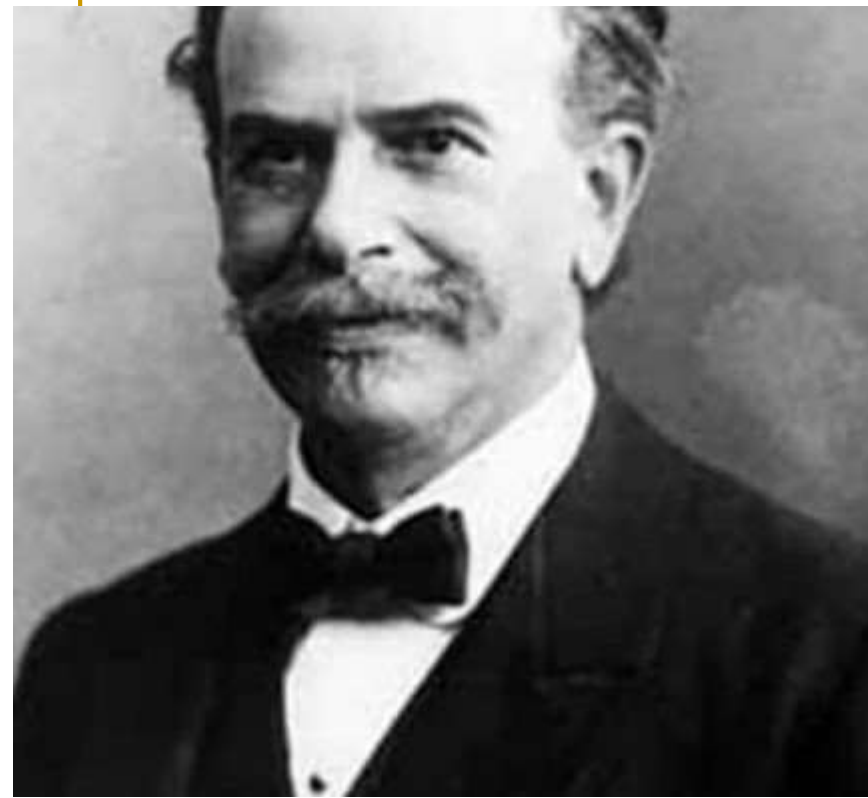
My personal hero



A bit of history

- Edmond Edmont 1896-1900 surveyed 639 rural locations in French-speaking areas of France, Belgium, Switzerland and Italy using 1900 item questionnaire
 - for Jules Gillieron's *Atlas Linguistique de la France* (published in 13 volumes 1902-1910)
 - Became a model for dialectology data collection elsewhere, not seriously challenged until 1960s
-

Our next hero – Papa Franz



Read more in King 2019



Language documentation 1

- Term widely used in late 19th and early 20th century to refer to the study of indigenous languages in the Boasian tradition, characterised by:
 - brief summer fieldwork
 - collection of dictated texts, vocabulary and grammatical forms
 - part of broad anthropological enterprise to ‘save’ disappearing cultures
 - part of a humanistic enterprise to understand the nature of human beings and societies, combatting racism and discrimination (King 2019)
 - training and engagement of native speakers as data producers and co-authors
 - use of latest technology



- goal: production of 'Boasian trilogy': text collection, grammar, dictionary
- (much material ends up in archives but not as a goal)

Language documentation 2

- “concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties” (Himmelman 1998)
- Features:
 - *Focus on primary data and analysis*
 - *Accountability*
 - *Long-term storage and preservation of data and analysis*
 - *Interdisciplinary teams*
 - *Cooperation with and direct involvement of the speech community*

Language documentation – outcomes

- *Narrow view*: outcome is **annotated and translated corpus** of archived representative materials on use of a language, cf. DoBeS/TLA, ELAR – separate from **description** (language as system)
 - *Broad view*: outcome is transparent records of a language (“for philologists in 500 years time”), with description and theorisation dependent on them (Woodbury)
-

McGill Cicipu corpus

Elan - svtmg001.eaf

File Edit Annotation Tier Type Search View Options Window Help

Grid Text Subtitles Audio Recognizer Metadata Controls

ft@Tenii

Nr	Annotation	Begin Time	End Time	Duration
68	yes	00:03:27.030	00:03:27.790	00:00:00.760
69	Hausa has swept them away	00:03:28.260	00:03:29.450	00:00:01.190
70	yes	00:03:30.260	00:03:30.960	00:00:00.700
71	if they want proper Cicipu, it's only the former people	00:03:33.910	00:03:37.830	00:00:03.920
72	m-hm, not the children of now	00:03:38.350	00:03:40.310	00:00:01.960
73	Cicipu only the former people	00:03:40.310	00:03:42.310	00:00:02.000
74	yes	00:03:42.310	00:03:44.310	00:00:02.000
75	yes	00:03:46.550	00:03:47.570	00:00:01.020
76	yes	00:03:58.230	00:03:59.010	00:00:00.780
77	one place, doing their work in one place	00:04:00.550	00:04:03.790	00:00:03.240
78	they would do it together	00:04:06.070	00:04:07.870	00:00:01.800

00:03:29.710 Selection: 00:00:00.410 - 00:00:01.490 1080

Selection Mode Loop Mode

100 00:03:26.000 00:03:26.500 00:03:27.000 00:03:27.500 00:03:28.000 00:03:28.500 00:03:29.000 00:03:29.500 00:03:30.000 00:03:30.500 00:03:31.000

[117] 196 svtmg001.098 svtmg001.099 svtmg001.100

ref@Tenii [197] óo tìkógó tìhúdò rè óo

tx@Tenii [671] yes Hausa has swept them away yes

ft@Tenii [197]

ref@Yaaki [12] svtmg001.097

tx@Yaaki [62] húuw tì áná tìkógó kóo?

ft@Yaaki [12] they speak it like Hausa?

Cicipu annotations

Toolbox - Dictionary.txt

File Edit Database Project Tools Checks View Window Help

[no filter]

Texts.txt

Reference	svtmg001.099				
Start Time	208.260				
End Time	209.450				
Speaker	Tenii				
Text	tùkógó tìhúfò rè				
Morphemes	tì- kógó ti- hufò -L -H -L rè				
Gloss	NC6- Islam AGR6- sweep -RLSp 3PP.PRO				
Part of Speech	nc- n agv- v -vtonep pro				
Free Translation	Hausa has swept them away				
Reference	svtmg001.100				
Start Time	210.260				
End Time	210.960				
Speaker	Tenii				
Text	óo				
Morphemes	ò'fi				
Gloss	yes				
Part of Speech	intj				

Dictionary.txt

\lx Lexeme	\lc Citation form	\ge English Gloss	\ps Part of speech	\gn Gloss (n)	\pdv Paradigm for
kógó	*empty*	Hausa_person	n	*empty*	8/2
kókó	(mò)-kókó	small_drum_k.o.	n	*empty*	4?/5?
kókóp	*empty*	drum_k.o.	n	kuge?	8/3?
kómmó	(kù)-kómmó	dirt?	n	*empty*	9/2
kómó	*empty*	cover	v	*empty*	*empty*
kómó	(ù)-kómó	salt	n	*empty*	7/8
kóó	*empty*	early	ideo	da wuri	*empty*
kóó	(kà)-kóó	egg	n	*empty*	1/2
kóó	(ù)-kóó	floor	n	farfajiya	7
kòòhúu	*empty*	lung	n	huhu ; kuhu	8
kóré	*empty*	indeed	adv	kwarai	*empty*
kótí	(mò)-kótí	stump	n	*empty*	4/5
koto	*empty*	finish	v	karke	*empty*
k-átà	(k-à)-k-átà	back of head	n	*empty*	1/2

Metadata.txt

\id	\ti title	\con contributor	\af audio file	\p recording place
svgd001	Discussion of chieftanc	GDM ; JN ; King	..\..\audio_visual\svgd001.wav	King's guest house, Korisino
svmk001	Greeting the Malhu	MM ; MK	..\..\audio_visual\svmk001.wav	Mallu's guest hut in Kadaada
svmy001	Norman Biggs' grave	MY	..\..\audio_visual\svmy001.wav	At the grave of Norman Biggs, Sakaba
svsdt001	Ukula mountain	MoMu ; SDT (Sani the	..\..\audio_visual\svsdt001.wav	On the Ukula (Maburya) mountain
svtmg001	Interview about the old	MoMa ; TMG ; Yaaki (..\..\audio_visual\svtmg001.wav	TMG's compound in Ka'ingawa KaGaladima

\lx kógó 836/1950 Cicipu.prj

Cicipu archival deposit

Cicipu documentation

[Home](#) [Resources](#)

Search this deposit

[Reset keywords](#)

Access protocol

U R C I S (407)

Language [more ▼](#)

Cicipu (259)
Damakawa (2)
Duka (2)
English (1)
Hausa (8)
[more ...](#)

Type

Audio (287)
Document (20)
ELAN (60)
Image (141)
Text (37)
Transcriber (32)
Video (53)
XML (2)
Zipped collection (7)

Tags

Kezzeme (5)
Photo (1)
Photos (8)

Genre [more ▼](#)

Christian (2)

Cicipu documentation

Language: Cicipu [awc]

Depositor: Stuart McGill

Location: Nigeria

Summary of deposit

This corpus contains folktales, riddles, historical narratives, casual conversation, commentaries on festival videos, interviews, songs, prayers, and sermons. Nine Pear Film narratives are also included. In total there are approximately six hours of interlinearised time-aligned texts are provided in Toolbox/ELAN format. The corpus also contains an accompanying lexicon in Toolbox format, collected from the texts as well as from the SIL Africa Area 1700-item wordlist. A large number of elicitation sessions are also provided (conducted in either Hausa or Cicipu). GPS data of the Cicipu area is included.

Group represented

The Acipu of Kebbi and Niger State, Nigeria

Language information

Called Acipanci in Hausa. Called 'Western Acipa' in Ethnologue 15th edition.

Special Characteristics

The deposit includes A. B. Mathews' 'Historical and anthropological report on the Achifawa', an unpublished typewritten manuscript from 1926. There is a physical copy in the National Archives, Kaduna (K2, 068), from which the electronic copy in this corpus was photographed.

Deposit status

✓ **Curated:**
Resources online and curated

Depositor

Stuart McGill



Nationality: UK

Affiliation: School of Oriental and African Studies



Language documentation 2 – drivers

- developed since 1995 in response to the urgent need perceived by researchers to make an enduring record of the world's many endangered languages and to support speakers of these languages in their desire to maintain them, fuelled also by developments in information, media, and communication technologies
- concerned with roles of language speakers and communities and their rights and needs
- is not limited to endangered languages – can be applied to any linguistic variety with any level of vitality

What's new in language documentation 2?

- **Data focus** – Himmelman's "primary data", but also **structured data** derived from processed materials (transcribed, translated, annotated digital files). A collection of such material is called a **corpus**. See Himmelman 2012.
- **Accountability** – we expect the materials ("primary" and analysed) to be made available to others. Some have argued for **reproducibility**, i.e. the possibility of recreating the researcher's analytical steps to see if the outcome is the same (or different). See Berez-Kroeker et al 2017. We discuss this later.
- **Preservation** – long-term storage in safe archival facilities where the data and analysis (corpora) can be safeguarded for the long term (including refreshing data formats to take into account changing software)
- **Reliance on software tools** – data and analysis is stored in digital files and access is mediated via computer software

Language description

- Looks at language as a structural system, abstracted away from use
 - Is concerned with questions like:
 - What is a language system/grammar?
 - To what extent are languages alike and to what extent are they different?
 - What does this tell us about the human mind?
 - What does this tell us about human communication?
 - How does a language system work and how is it acquired?
-

Language description requires

- Asking the right questions/collecting relevant data. Rice (2005: 236) argues that formal syntactic theory forces a grammar writer to ask questions that are not very likely to be asked otherwise.
- Making generalisations and drawing distinctions about the grammar of languages. In other words, descriptions must be generalizable, rather than simply observational, i.e., must represent broad statements about the described linguistic system.
- Labelling and categorizing the phenomena in one way or another (i.e., you need a 'metalanguage', comparative concepts, terminology ...)

Language description requires

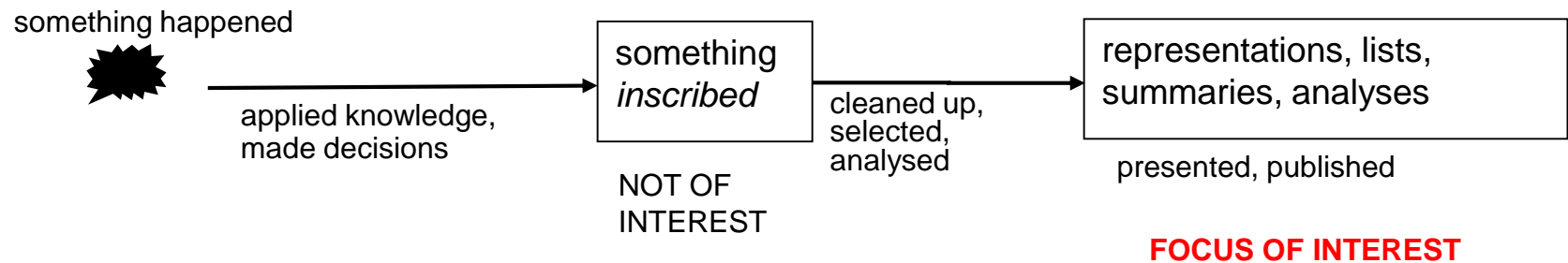
- A theory ('framework') that underlies the labels and categories, e.g., 'generative' or 'functional' mechanisms, and a model for argumentation and explanation
 - Presenting data and analyses in a way that is acceptable and interesting to a wider audience – a “grammar” or “dictionary” as an academic object, organized in a particular socio-culturally accepted way
-

Documentation <--> Description

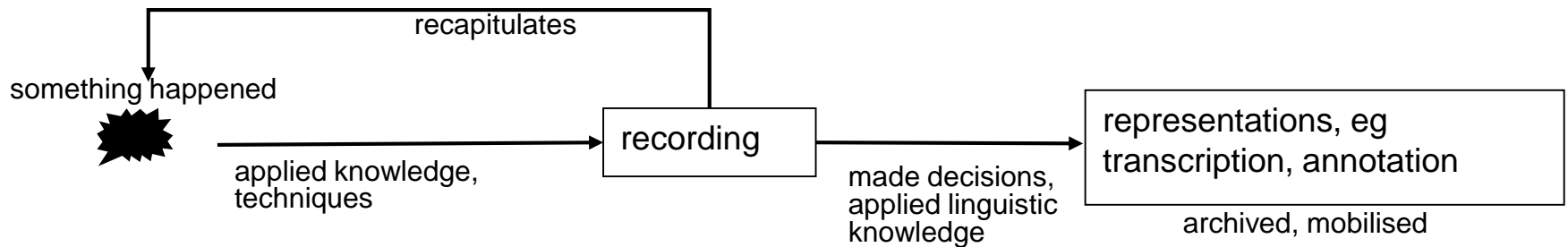
- Himmelmann 1998 claims they are essentially separate activities and have different epistemologies, methods, and goals
- Description typically uses a narrower range of methods than language description: elicitation (word lists, questionnaires, translation, grammaticality judgements) vs. participant observation and data collection in its socio-cultural context ('naturalistic language', e.g. conversation), and/or experimentation (stimuli, games).
- Descriptive sources often not tracked (Gawne et al. 2017) and hence research is not reproducible (Berez-Kroeker et al. 2019)

Workflow

Description



Documentation



FOCUS OF INTEREST

FOCUS OF INTEREST

Description vs. Documentation 2

- Documentation 2 needs an epistemology for media capture – audio and video recording
- Need to pay attention for good practices in recording – eg. microphone choice and spatiality in audio, framing-lighting-editing for video (“recording arts”)
- Some concern for socio-cultural context (‘ethnography of speaking’)
- Concern for data structuring and data management – eg. ‘portability’, relational modelling, XML
- Concern for ‘standards’ and cross-project comparability, especially typology and data mining
- Concern for ethics of research – documentation collects language use in ‘intimate’ personal contexts, impacts on potential users and uses of documented speech events
- Changing models of research and relationships with people

Components of documentation

- *Planning* – language, funding, fieldwork, equipment
- *Recording* – of media and text (including metadata) in context
- *Transfer* – to data management environment
- *Adding value* – transcription, translation, annotation, notation and linking of metadata
- *Archiving* – creating archival objects, assigning access and usage rights
- *Mobilisation* – creation, publication and distribution of outputs

Recording

- *audio* – basic and familiar in modern linguistic work.
Important considerations: environment, equipment choice, microphones, monitoring, file type (wav not mp3 generally recommended)
- *video* – immediate, rich in authenticity, multi-dimensional in context, great interest to communities, can be produced independently by community members BUT more difficult to produce, process, access without time-aligned annotation, transfer, store and preserve
- *text* – compact, stable, easy to store, access and index, can express hypertextual links to other text and media BUT relies on literacy and is less rich than audio/video

□ *metadata* – data about the data: needed to identify, manage, retrieve data. Provides context and understanding of data to oneself and others. Types:

- Cataloguing — identifying and locating data, eg. language code, file id, recorder, speaker, place of recording, date of recording etc
 - Descriptive — kind of data found in a file, eg. abstract/summary of file contents, knowledge domain represented
 - Structural — specification of file organisation, eg. textfile is a bilingual dictionary
 - Technical — file format, kind of software needed to view, preservation data
 - Administrative — work log, intellectual property rights, moral rights, access and distribution restrictions
-

-
- ❑ *meta-documentation* – documentation of language documentation models, processes and outcomes, goals, methods and conditions (linguistic, social, physical, technical, historical, biographical) under which the data and analysis was produced (should be *as rich and appropriate* as the documentary materials themselves)
-

Adding value

- requires decision making (selection, editing, choice of method and theory) and is very time consuming (eg. annotation can be 100:1 in terms of time required)
- linguistic value adding ('thick' meta-data):
 - *transcription* – textual representation of audio signal (orthographic, phonemic, phonetic) typically time-aligned to media
 - *annotation* – overview, code, morphological, grammatical, semantic ('gloss'), syntactic, pragmatic, discourse. Fixation among documenters on 'interlinear glossing', cf. overview annotation/summary
 - *translation* – levels: word, sentence, paragraph, text. Types: literal, running, parallel, literary (Woodbury 2005, Evans & Sasse 2005)

Tools for value adding

- application programs, components, fonts, style sheets
 - application programs:
 - *general purpose* software – user must design data structures and manipulation routines, eg. LibreOffice, MS Office (Word, Excel, Access)
 - *specific purpose* software – designed for particular tasks, eg. Transcriber, ELAN, Arbil, FLEx, Toolbox, SayMore
 - Important: design and use a workflow that enables data transfer (export, import) without loss/corruption of encoded knowledge
-

Archiving

A digital language archive:

- is a trusted repository created and maintained by an institution with a commitment to the long-term preservation of archived material
 - has policies and processes for acquiring, cataloguing, preserving, disseminating, and format/content migration
 - is a platform for building and supporting relationships between data providers and data users
-

Mobilisation

- Creation of usable outputs for a range of different audiences, eg. multimedia websites, sub-titled video, apps
- There are tools to help with this (LexiquePro, CuPed) and people working on app development who can help, e.g. Ma! Iwaidja



Frameworks for language research



Ethical
research

Advocacy
research

Collaborative
research

Empowering
research

(Cameron, Frazer, Harvey, Rampton, and Richardson 1992)

1. “Research ON a language”

- Usual in first half of 20th century
 - “Salvage linguistics”
- Who is language documentation for?
- Still continues: ‘lone wolf’ linguist encouraged by some funding models
- “Community members report sometimes feeling that the linguist comes in, reifies the language, turns it into a commodity, and then takes it away.” (Bower 2011: 468)

2. “Research FOR the community”

- Developed in 1960s
 - period of civil rights movements in USA
- Fieldworkers ‘give something back to the community’
 - e.g. educational materials,
 - advocacy: Labov 1982
- Endangered language speakers are not just sources of data
 - often economic and social problems contribute to language shift
- Not all linguists have other needed skills (e.g. social work, medical expertise)

3. “Research WITH the community”

- Developed in 1980s
 - “Action Research”
 - “Negotiated fieldwork”
 - Equal say and partnership to speakers of the language
 - Full participation, from planning to outputs
 - Now dominant model
 - at least in rhetoric!
 - May be difficult to find funding for
-

4. “Research BY a community”

- The project is community-driven
 - May include maintenance/revitalisation measures, creating language teaching programmes, etc.
 - e.g. Dieri Aboriginal Corporation
 - Multidisciplinary approach
 - Role of external linguist:
 - Training, teaching, mentoring native speakers ...
-

Issues in language documentation 2

- Objectification and commodification of languages
 - ‘Community members report sometimes feeling that the linguist comes in, reifies the language, turns it into a commodity, and then takes it away.’ (Bower 2011: 468)
 - ‘Technical parameters such as bit rates and file formats are now often foregrounded to the point that they eclipse discussions of documentation methods’ (Dobrin, Austin & Nathan 2009: 42)
- Arguably, we should document language ecologies, not just individual languages
 - Multilingual repertoires, mixed codes, translanguaging, contact effects (Mühlhäusler 2003, Grenoble 2011)

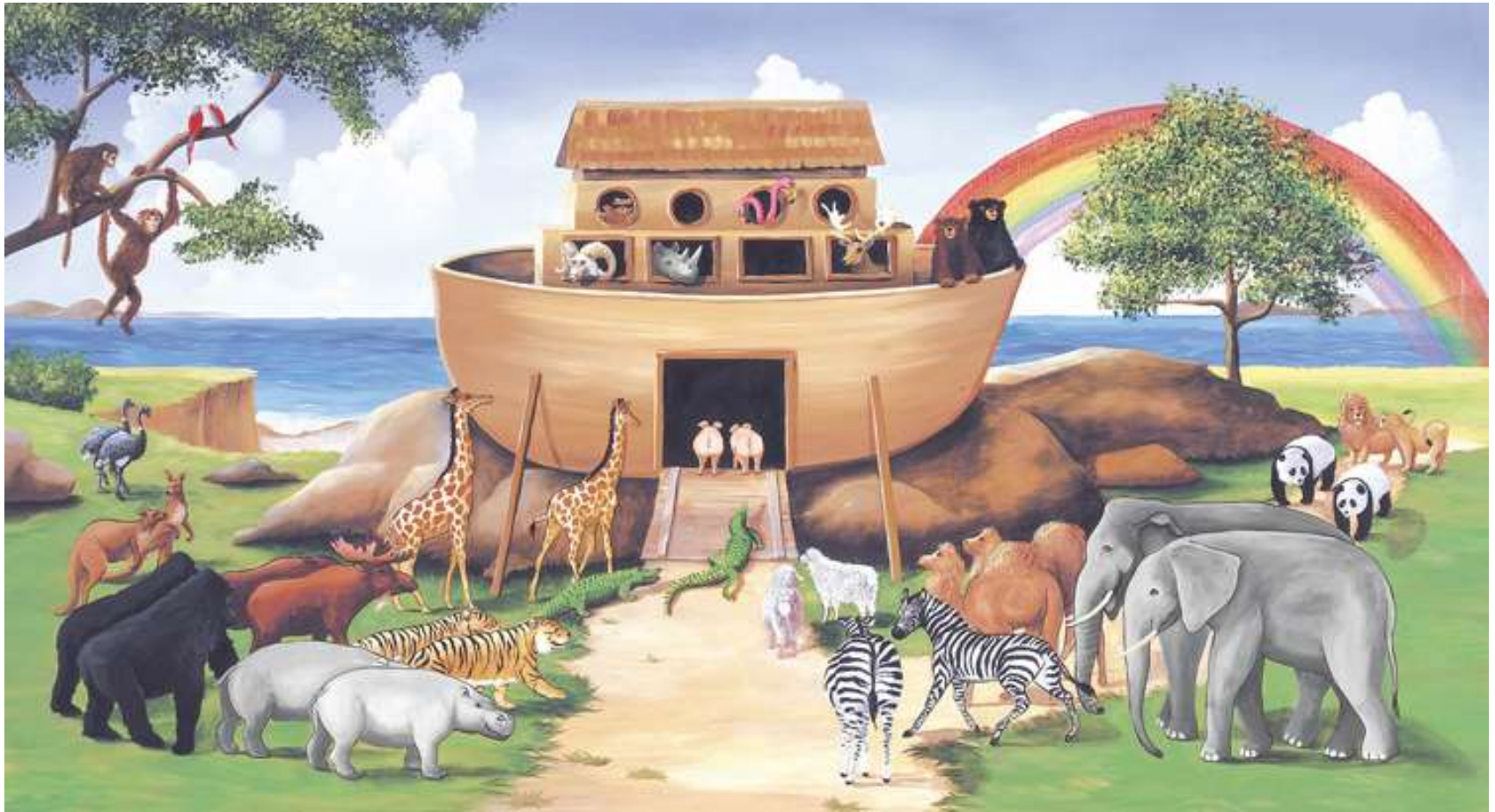
-
- **lack of audio skills:** little or no knowledge about recording arts and microphone types, properties and placement (microphone choice and handling is the single greatest determiner of recording quality)
 - **video madness:** video recordings made without reference to hypotheses, goals, or methodology, simply because the technology is available, portable and relatively inexpensive
 - **corpus taming:** little ability at corpus and metadata management, file naming and bundle organisation – ELAR spent huge amounts of time and energy simply cleaning up deposits before they could be archived.
-

Despite the rhetoric



- lone wolf linguists primarily focussed on language
- little real interdisciplinary interest
- the linguist decides what to deliver to academia and communities and produces familiar and traditional outcomes (dictionaries, orthographies, story collections, etc.)

The documentation model 1995-2010



Noah's arc(hive) – saving the morphemes 2-by-2

There is an output gap



The output gap

Outputs from language documentation projects have bifurcated into:

- ❑ **Published** grammars, (bilingual) dictionaries and (glossed) texts – ‘revival’ of familiar genres linguists have been comfortable with for 100+ years
- ❑ **Archive** deposits – hundreds or thousands of files, professionally curated by archivists, but often poorly organised or structured, with little if any contextualisation

Corpus accessibility – I found it, what now?

Cicipu documentation

Home Resources

Found 60 bundles in this deposit with keyword **ELAN x** (page 1 of 8)


1 2 3 4 5 6 7 8 next > last »

▼ **Discussion of chieftancy**

svgd001.eaf Access protocol: **URCS**

[Download](#)

svgd001.001.mpg Access protocol: **URCS**



00:05 00:17

[Download](#)

Search this deposit

[Search](#)

[Reset keywords](#)

Access protocol

URCS (60)

Language

[more](#) ▼

- Cicipu (58)
- Tidipo (5)
- Tikula (3)
- Tirisino (6)
- Damakawa
- [more...](#)

Type

ELAN x

- Audio (60)
- Image (8)
- Transcriber (3)
- Video (10)
- Document
- Text
- XML
- [Zipped collection](#)

Tags

- Kezzeme (2)
- Photo
- Photos

Genre


[more](#) ▼

Deposit status

✓ **Curated:**
Resources online and curated

Depositor

Stuart McGill




Nationality: UK
Affiliation: School of Oriental and African Studies

Your access

Your roles: **URCS**

Tools

[Download metadata](#)
[Add to My Bookmarks](#)



Corpus accessibility – I can't even find it

The screenshot displays the 'The Language Archive' website interface. On the left is a hierarchical tree view of the archive's contents, including categories like 'IMDI-corpora', 'AILLA', 'ANDES', 'Bavarian Archive for Speech Signals (BAS)', 'CLARIN NL', 'CORP-ORAL', 'DBD', 'DoBeS archive', and various language-specific projects. The main panel on the right shows the details for a specific session, 'DJI1010312CDD'. The session details include a title 'Tree list', a date '2012-03-01', and a description 'Verification of the pronunciation and agreement patterns of all tree names'. Below this, there are sections for 'Location', 'Project', 'Content', 'Languages', 'Actors', and 'MediaFile'. The 'Content' section lists various metadata fields such as Genre, SubGenre, Task, Modalities, Subject, Interactivity, PlanningType, Involvement, SocialContext, EventStructure, and Channel. The 'Languages' section lists 'Language Bainounk Gubeeher (c)' and 'Language French (c)'. The 'Actors' section lists 'Actor Alexander Cobbinah' and 'Actor Jean Marie Sagna'. The 'MediaFile' section lists 'Type audio', 'Format audio/x-wav', 'Size 460 MB', and 'Quality Unspecified'. The 'RecordingConditions' section lists 'Start Unspecified' and 'End Unspecified'.

The Language Archive

about manual register user: anonymous Log in

METADATA SEARCH CONTENT SEARCH MANAGE ACCESS REQUEST ACCESS

CITATION DOWNLOAD ALL VERSION INFO

Session

Name DJI1010312CDD
Title Tree list
Date 2012-03-01

Description

Verification of the pronunciation and agreement patterns of all tree names.

Location

Project DoBeS 3P

Content

Genre Elicitation
SubGenre lexical elicitation
Task
Modalities speech
Subject
Interactivity interactive
PlanningType planned
Involvement
SocialContext
EventStructure
Channel

Languages

Language Bainounk Gubeeher (c)
Language French (c)

Actors

Actor Alexander Cobbinah
Actor Jean Marie Sagna

MediaFile

Type audio
Format audio/x-wav
Size 460 MB
Quality Unspecified

RecordingConditions

TimePosition

Start Unspecified
End Unspecified

What is missing?

- Meta-documentation – the documentation of documentation projects, goals, methods, IP contributions, outcomes
- New (unfamiliar) genres that link and contextualise analytical outputs and the archival corpus:
 - ethnographies of documentation project designs
 - accounts of data collection (cf. archaeology ‘field report’)
 - finding-aids to corpus collections
 - ‘exhibitions’ or ‘guided tours’ of archival deposits
- Evaluation measures that enable properly-based peer assessment of documentations, equivalent to the way traditional outputs are judged

Language Documentation – future?

Diversity

of goals, contexts, people, data, corpora, outcomes

- ❑ move away from Noah's Arc(hive) to more focused documentation, with increased participant observation, eg. ELDP 2012 grant list: bark cloth making, libation rituals, fishing practices, child language, interactive speech, and ethnobotany
 - ❑ diverse **inputs** – field interviews, experiments and observations (traditionally the bread and butter of documentation and description) but also Youtube uploads, Twitter feeds, Facebook, blogs, email, chat, Skype, local pedagogy in revitalisation
 - ❑ diverse **outputs** – books, papers and archive deposits (the bread and butter of 1990's documentation) but also Youtube uploads, Twitter posts, Facebook, blogs, email, chat, Skype, local pedagogy in revitalisation, mobile apps, Kindle readers
-

New genres

- Woodbury (2015) ‘Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire’:

“I urge documenters to take **authorial control** of their work, as they would if each archived collection were a book of language materials

- make a guide to your own documentary corpus
- include meta-documentation: describe the design of activities or projects from which the corpus arose, offer a theorization of the corpus (or several, from different perspectives), and describe the appraisal process used to select and assemble the corpus
- write narratives, logs, and journals
- think of your corpus as belonging to a **genre**.

To some extent, all this means documenters taking on some of the work traditionally done by archivists.”

Transdisciplinarity

- Is language documentation a new sub-field of linguistics? (as per Himmelmann, Austin) or
 - Is it a new transdisciplinary approach that: “must draw on concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts and more” (Woodbury 2011), where “more” includes history, archiving, museum studies, project management, creative writing, social media, ornithology, biology (cf. PAW project at SOAS), political science, development studies?
-

Transdisciplinarity

- Siebert (2016) ‘Documentary linguistics: a language philology of the 20th century’:

“documentary linguistics’ focus on ‘direct representation of discourse’ requires a broader conceptualization of the field that moves **beyond purely linguistic concerns**. This article recasts documentary linguistics as a philology, broadly understood as the inquiry into ‘the multifaceted study of texts, languages, and the phenomenon of language itself’ ... The reconceptualization of documentary linguistics described in this article opens documentary linguistics to non-core linguistic types of language documentation efforts and situates the documentary activities more broadly in the humanistic enterprise of communicating, discussing, studying, and understanding human achievements of other times and places.”

Conclusions

- Some researchers have presented language documentation as a return to the Boasian past while others see it as a new approach to the study of human language that pays better attention to data collection and analysis, and to communities, contexts and roles
- it appeared to be an opportunity to shift the socio-political academic balance between “fieldworkers” and “armchair linguists” (typologists, theoreticians) by providing a foundation (theory, best practices) for documentation, in contrast to language description
- Over the past 20 years, and especially the last 10 years, we have seen shifts in the goals, methods, foci and contexts of Language Documentation to make it more pluralistic, open, and socially networked and responsive
- However challenges remain, including encouraging new genres that bridge the output gap, more reflexivity, and better engagement with transdisciplinarity and the ethnography of our research and its contexts

Thank you!

References

- Austin, Peter K. 2013. Language documentation and meta-documentation. In Mari Jones & Sarah Ogilvie (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, 3-15. Cambridge: Cambridge University Press.
- Austin, Peter K. 2016. Language documentation 20 years on. In Martin Pütz & Luna Filipović (eds.) *Endangered Languages and Languages in Danger: Issues of ecology, policy and human rights*, 147-170. Amsterdam: John Benjamins.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2017. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1), 1-18.
-

References

Childs, Tucker, Jeff Good & Alice Mitchell. 2014. Beyond the ancestral code: Towards a model for sociolinguistic language documentation. *Language Documentation and Conservation* 8, 168–191.

Dobrin, Lise, Peter K. Austin & David Nathan. 2009. Dying to be counted: the commodification of endangered languages in language documentation. *Language Documentation and Description* 6, 37-52.

Dobrin, Lise & Saul Schwartz. 2020. The social lives of linguistic field materials. To appear in *Language Documentation and Description* 20.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11, 157–189.

References

- Grenoble, Lenore. 2010. Language documentation and field linguistics: The state of the field. In Grenoble, Lenore A. and N. Louanna Furbee (eds.) *Language Documentation: Practice and values*, 289-309. Amsterdam: John Benjamins Publishing Company.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36, 161–195.
- Himmelman, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation and Conservation* 6, 187-207.
- King, Charles. 2020. *Gods of the Upper Air*. New York: Anchor Books.
- Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Toshihide Nakayama & Keren Rice (eds.) *The Art and Practice of Grammar Writing*, 135-157. Language Documentation & Conservation Special Publication No. 8.
-

References

Rice, Keren. 2006. Let the language tell the story? The role of linguistic theory in writing grammars. In Felix K. Ameka, Alan Charles Dench & Nicholas Evans (eds.) *Catching Language: The Standing Challenge of Grammar Writing*, 235-268. Berlin: Mouton de Gruyter.

Seidel, Frank. 2016. Documentary linguistics: A language philology of the 21st century. *Language Documentation and Description* 13, 23-63.

Wilbur, Joshua. 2014. Archiving for the community: Engaging local archives in language documentation projects. *Language Documentation and Description* 12, 85-102.

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation & Description* 1, 35-51.

References

Woodbury, Anthony C. 2011. Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 159-186. Cambridge: Cambridge University Press.

Anthony C. Woodbury (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description* 12, 19-36
