

**REFLECTIONS
ON LANGUAGE
DOCUMENTATION
20 YEARS AFTER
HIMMELMANN 1998**

edited by
Bradley McDonnell
Andrea L. Berez-Kroeker
Gary Holton

Language Documentation & Conservation
Special Publication **15**

Reflections on Language Documentation
20 Years after Himmelmann 1998

edited by

Bradley McDonnell
Andrea L. Berez-Kroeker
Gary Holton

Published as a Special Publication of Language Documentation & Conservation

Language Documentation & Conservation
Department of Linguistics
University of Hawai'i at Mānoa
Moore Hall 569
1890 East-West Road
Honolulu, Hawai'i 96822
USA

<http://nflrc.hawaii.edu/ldc>

University of Hawai'i Press
2840 Kolowalu Street
Honolulu, Hawai'i
96822-1888
USA

© All texts and images are copyright to the respective authors, 2018

© All chapters are licensed under Creative Commons Licences

Cover design by Jack DeBartolo 3

Library of Congress Cataloging in Publication Data

ISBN: 978-0-9973295-3-7

<http://hdl.handle.net/10125/24800>

Contents

1 Introduction	1
<i>Bradley McDonnell, Gary Holton & Andrea L. Berez-Kroeker</i>	
RE-IMAGINING DOCUMENTARY LINGUISTICS	
2 Reflections on the scope of language documentation	13
<i>Jeff Good</i>	
3 Reflections on reproducible research	22
<i>Lauren Gawne & Andrea L. Berez-Kroeker</i>	
4 Meeting the transcription challenge.	33
<i>Nikolaus P. Himmelmann</i>	
5 Why cultural meanings matter in endangered language research	41
<i>Lise M. Dobrin & Mark A. Sicoli</i>	
6 Reflections on (de)colonialism in language documentation.	55
<i>Wesley Y. Leonard</i>	
7 Reflections on public awareness	66
<i>Mary S. Linn</i>	
KEY ISSUES IN LANGUAGE DOCUMENTATION	
8 Interdisciplinary research in language documentation	76
<i>Susan D. Penfield</i>	
9 Reflections on language community training.	86
<i>Colleen M. Fitzgerald</i>	
10 Reflections on funding to support documentary linguistics	100
<i>Gary Holton & Mandana Seyfeddinipur</i>	
11 Reflections on ethics: Re-humanizing linguistics, building relationships across difference.	110
<i>Ewa Czaykowska-Higgins</i>	
12 Reflections on diversity linguistics: Language inventories and atlases. . .	122
<i>Sebastian Drude</i>	
13 Reflections on the diversity of participation in language documentation	132
<i>I Wayan Arka</i>	

14 Reflections on software and technology for language documentation . . .	140
<i>Alexandre Arkhipov & Nick Thieberger</i>	

BEYOND DESCRIPTION: CREATING & UTILIZING DOCUMENTATIONS

15 Reflections on descriptive and documentary adequacy	151
<i>Sonja Riesberg</i>	
16 Reflections on documentary corpora	157
<i>Sally Rice</i>	
17 Reflections on the role of language documentations in linguistic research	173
<i>Stefan Schnell</i>	
18 Reflections on documenting the lexicon	183
<i>Keren Rice</i>	
19 Reflections on linguistic analysis in documentary linguistics	191
<i>Bradley McDonnell</i>	

VIEWS ON LANGUAGE DOCUMENTATION FROM AROUND THE WORLD

20 Reflections on linguistic fieldwork	202
<i>Clarie Bower</i>	
21 The state of documentation of Kalahari Basin languages	210
<i>Tom Güldemann</i>	
22 From comparative descriptive linguistic fieldwork to documentary linguistic fieldwork in Ghana	224
<i>Felix K. Ameka</i>	
23 Caucasus — the mountain of languages	240
<i>Manana Tandashvili</i>	
24 Reflections on language documentation in India	248
<i>Shobhana Chelliah</i>	
25 Reflections on linguistic fieldwork and language documentation in eastern Indonesia	256
<i>Yusuf Sawaki & I Wayan Arka</i>	
26 Reflections on linguistic fieldwork in Australia	267
<i>Ruth Singer</i>	
27 In search of island treasures: Language documentation in the Pacific . . .	276
<i>Alexandre François</i>	
28 Reflections on language documentation in the Southern Cone	295
<i>Fernando Zúñiga & Marisa Malvestitti</i>	
29 Reflections on language documentation in the Chaco.	303
<i>Lucía Golluscio & Alejandra Vidal</i>	
30 Reflections on fieldwork: A view from Amazonia.	321
<i>Christine Beier & Patience Epps</i>	

31 Reflections on linguistic fieldwork in Mexico and Central America	330
<i>Gabriela Pérez Báez</i>	
32 Reflections on language documentation in North America	340
<i>Daisy Rosenblum & Andrea L. Berez-Kroeker</i>	

1

Language Documentation & Conservation Special Publication No. 15
Reflections on Language Documentation 20 Years after Himmelmann 1998
ed. by Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton, pp. 1–11
<http://nflrc.hawaii.edu/l dc/>
<http://hdl.handle.net/10125/10125/24803>

Introduction

Bradley McDonnell
Gary Holton
Andrea L. Berez-Kroeker
University of Hawai‘i at Mānoa

This chapter introduces the volume, *Reflections on Language Documentation 20 Years after Himmelmann 1998*, providing a short justification for the volume, summarizing each of the four major parts of the volume, and identifying major themes that emerge in the 31 chapters. It concludes by noting some of the volume’s limitations.

1. Introduction¹ Twenty years ago Himmelmann (1998) envisaged a radical (new?) approach to the science of language, recognizing a fundamental distinction between documentation and description and focusing on the collection of primary data which could be repurposed and serve as an enduring record of endangered languages. (For a concise overview of the development of language documentation over the past 20 years, see Austin 2016.) Of course Himmelmann was not the first to argue for an approach to linguistics grounded in data, especially the collection of text corpora. Some of the earliest field workers in the modern era—from Franz Boas to Edward Sapir to P.E. Goddard to Melville Jacobs—all saw the value of primary text collections and recordings (cf. Boas 1917). And the call to arms for renewed focus on endangered language documentation had been issued well before Himmelmann’s seminal paper was published (cf. Krauss 1992). However, by carefully articulating the distinction between documentation and description, Himmelmann (1998) clarified a truth that had been hidden within the everyday work of linguistics. It is a truth not far from the mind of every descriptive linguist, but one which had not often been discussed.

Most, if not all, linguistic field workers engage in both documentation and description, collecting and annotating primary recordings but then also analyzing those data to

¹We would like to thank the contributors to this volume for their time and effort in not only writing thoughtful reflections, but also for their timely reviewing. We also thank Nikolaus Himmelmann for early conversations about the direction this publication should take. We also thank participants in the special session “20 Years of Language Documentation” at the 22nd Foundation for Endangered Languages Conference in Reykjavik for fruitful early feedback on the volume.

extract a description couched in meta-linguistic terms. Some authors have objected to Himmelmann (1998) precisely because these two activities—documentation and description—are often inseparable in practice. Yet, in our opinion, this objection misses the point. While the two activities may go hand in hand, the products suffer very different fates, with the former much more likely to have a lasting impact on the field. Woodbury (2011) defines documentary linguistics as “the creation, annotation, preservation and dissemination of transparent records of a language” (159). As anyone consulting a descriptive grammar written in an obscure syntactic framework can attest, linguistic description generally fails to meet the transparency requirement (cf. Gawne et al. 2017).

The recognition that documentation, not description, is more likely to have a lasting impact requires us to rethink the way we do linguistics. This does not mean that we need to stop doing description. It also does not give us license to ignore linguistic theory—in fact, quite the opposite. A renewed focus on primary data makes possible a data-driven science of linguistics in which theory is more robustly grounded in primary data, yielding reproducible results (cf. Berez-Kroeker et al. 2018). What this renewed focus does do is lead us to think more carefully about the collection of primary data as an end unto itself. This is the heart of the emerging field of documentary linguistics, as articulated by Himmelmann (1998). Now 20 years later we are able ask: how has this new field evolved?

In order to answer this question, we invited 38 experts from around the world to reflect on either particular issues within the realm of language documentation or particular regions where language documentation projects are being carried out. The issues addressed in this volume represent a broad and diverse set of topics from multiple perspectives and for multiple purposes that continue to be relevant to documentary linguists and language communities. Some topics have been hotly debated over the past two decades, while others have emerged more recently.

Thus, we asked each contributor to reflect on a particular issue in light of how the issue was originally raised and how it has developed in the field. While the original mandate was to reflect on the evolution of the field in the 20 years since Himmelmann (1998), many of the contributors have taken a longer view. In particular, many contributors speculate on what comes next, looking at the future of language documentation from a variety of perspectives. Hence, the 31 vignettes provide not only reflections on where we have been but also a glimpse of where the field might be headed. Based upon the subject matter, the chapters have been grouped into four parts:

Part 1: Re-imagining documentary linguistics These chapters re-imagine in some way how we conceive of documentary linguistics and how we carry out language documentations.

Part 2: Key issues in language documentation These chapters deal with key issues in language documentation that are already a part of the current discourse in the field, some of which have been discussed at length.

Part 3: Beyond description: Creating and utilizing documentations These chapters deal with the relationship between documentation and linguistic research and products thereof, such as dictionaries, grammatical descriptions, theoretical studies, and language corpora.

Part 4: Views on language documentation from around the world These chapters present a small sampling of language documentation (and linguistic fieldwork) in various regions around the world.

In the remainder of this introduction, we highlight some of the larger themes that have emerged in each of the four parts of the volume.

2. Part I: Re-imagining documentary linguistics The chapters in Part I take a broad view, envisioning a new future for documentary linguistics as a discipline. The authors of these chapters share the view that documentary linguistics should be imbued with a theoretical framework, not characterized as an atheoretical provider of data for linguistic analyses. To a large extent this view reflects a continuing maturing of the field. Documentary linguistics arose as a largely ad-hoc response to the endangered languages “crisis.” The race to catalog languages, develop recording and archiving standards, and implement funding schemes left little time to reflect on the scope of language documentation as a (sub)discipline of its own. Moving forward, Jeff Good (Chapter 2) suggests that we need to develop a documentary linguistics which is fully theorized and codified as a genuine subfield, taking care to consider aspects of languages which have often been omitted from the record. Good’s use of the label *documentary linguistics* as opposed to the generally synonymous term *language documentation* is deliberate, as the former suggests a genuine subfield of linguistics as opposed to an ancillary data-gathering activity.

Lacking a distinct theoretical framework, the field of language documentation has by default evolved with a focus on languages as the primary object of study. (Consider the huge effort still applied to the issue of distinguishing languages versus dialects.) This has happened in spite of Himmelmann’s stipulation that documentation should create a “record of the linguistic practices and traditions of a speech community” 1998: 166. A re-imagined, more fully theorized science of documentary linguistics could explore alternative approaches, for example by focusing on the linguistic behavior and knowledge of individuals rather than on particular speech events. Such an approach explicitly acknowledges the multilingual nature of speech communities which has been tacitly ignored by many documentation efforts

As several authors in Part I note, these new approaches to documentary linguistics have the potential to change the practice of linguistics more broadly. No where is this more apparent than in the discussion of Open Science and Reproducible Research. As Lauren Gawne and Andrea Berez-Kroeker (Chapter 3) note, the development of online digital archives has led to “a more open approach to data that would support research reproducibility” (p. 28). There is still much work to be done in order to fully realize the potential of open data in linguistics, but documentary linguistics is clearly leading the way toward creating a more data-driven science of linguistics, in which claims about language are grounded in data which can be accessed and assessed by other researchers. Looking ahead, Gawne & Berez-Kroeker call on linguists to value language documentation as fully as they have valued language description. Adopting and implementing professional standards such as the *Austin Principles of Data Citation in Linguistics* can help to move us in that direction.

Reflecting on the progress of documentary linguistics over the past twenty years, Nikolaus Himmelmann (Chapter 4) identifies a significant remaining challenge, which he labels the “transcription challenge.” This problem goes far beyond the practical issues of transcribing a large amount of collected data, i.e., the transcription bottleneck. Rather, the transcription challenge is about developing a “better understanding of the transcription process itself and its relevance for linguistic theory” (p. 35). Here Himmelmann echoes Good’s (Chapter 2) call for greater theorization of documentary linguistics. Transcription

is not a mechanical process but rather a creative act of “language making” which affects language ecology. Theorizing transcription requires an understanding of the cultural contexts of transcription, including speakers’ relationships to the normative aspects of writing (see also Dobrin & Sicoli, Chapter 5). Theorizing transcription also requires us to build language reclamation into the documentation cycle, as younger language transcribers may see themselves in the role of language learner through their work with older speakers (see also Leonard, Chapter 6). As Himmelmann concludes, “[i]t is only a minor exaggeration to say that language documentation is all about transcription” (p. 38, yet despite its important role in language documentation, transcription remains critically undertheorized and understudied. To the extent documentary linguistics has concerned itself with transcription, it has mostly focused on technical issues such as tools and standards. As Himmelmann reminds us, the conceptual separation of documentation and description provides an opportunity to focus on practices which have been overlooked or ignored in traditional descriptive linguistics. Given the key role of transcription in documentation, it clearly deserves a closer look.

Modern documentary linguistics was from the start envisioned as an interdisciplinary effort, yet as Lise Dobrin and Mark Sicoli (Chapter 5) point out, the anthropological perspective is often lacking. Even some of our most basic discourses about linguistic methodologies remain uninformed by community contexts. Who counts as a native speaker? How do speaker numbers get tabulated? Should language recordings be archived? None of these questions can be answered in the absolute but instead require an understanding of local cultural perspectives. For example, in situations of language shift, linguists may be tempted to define speakerhood in structural terms, whereas communities may place greater emphasis on command of cultural knowledge. Dobrin & Sicoli argue that—if local perspectives are taken into consideration—language documentation will be more successful to the extent that it recognizes the meanings that language has for local actors. Participant observation methodology provides a useful way to uncover these local meanings in language documentation work.

Although much of the rhetoric of documentary linguistics stresses the importance of community perspectives, the origins and practices of language documentation are often deeply rooted in colonial institutions which lie outside the control of local language communities. Intentionally or not, these power structures can sometimes work against Indigenous language communities. Wesley Leonard (Chapter 6) discusses some of the ways this occurs. Most notable is the false dichotomy between documentation and revitalization/reclamation—a distinction which is at best artificial and at worst detrimental to many language communities.

Leonard concludes his chapter by offering some potential “interventions” which can facilitate decolonial approaches to language documentation. Decolonial practices require rethinking the core values of linguistics in the academy, including such sacred cows as peer review. As Leonard notes, “[w]hen the language community is recognized as a core stakeholder, it follows that members of the language community will be among the reviewers of language documentation proposals and products” (p. 62). By recognizing speakers and speech communities as equal stakeholders in the documentation process—not as mere sources of data—we can not only counter the effects of colonialism but also improve the overall language documentation enterprise.

Mary Linn (Chapter 7) identifies an additional aspect of the effort to re-imagine documentary linguistics, namely, the need to engage with the wider public. As Linn argues, public awareness of the importance of language documentation is critical to

the success of documentation and conservation efforts and thus to the prevention of catastrophic language loss. Unfortunately, although endangered languages have featured prominently in the popular press over the past two decades, there is surprisingly little evidence of public support for language documentation and revitalization. Linn explores this topic via a qualitative review of public comments posted in response to articles about endangered languages published in major media outlets such as the BBC and the New York Times.

Language documentation obviously requires public support in the form of funding, be it directly through grants or indirectly through education. But public support is particularly critical to language conservation: if minority languages are to thrive, they will perforce do so alongside the majority languages which are currently threatening them. Engaging majority language communities in the appreciation of minority languages is thus critical to language survival. Unfortunately, changing public attitudes is notoriously difficult. Nevertheless, Linn sees one bright spot in the struggle to increase awareness and appreciation of endangered languages: namely, young people. She advocates for greater engagement with schools and teachers as a way to break down negative attitudes toward endangered languages. Her conclusion—“I have hope in change through youth” (p. 72)—signals an optimistic future for a re-imagined field of documentary linguistics.

3. Part II: Key issues in language documentation The chapters in Part II review some of the most important active discussions from the last two decades. Primary, perhaps, among these is the recognition of the need to “rehumanize” language documentation, to borrow a phrase from Ewa Czaykowska-Higgins (Chapter 11) and Dobrin & Berson (2011). That is, as a field we have done some serious reconsideration and reshaping of the archetype of the “lone wolf” linguist: models of fieldwork in which a single outsider enters a language community, gathers scientific data, and retreats back to the university are no longer acceptable. Instead, the goals of language documentation mandate us to include others. The inclusion of members of the language community in all aspects of the documentation project is a primary theme in chapters by I Wayan Arka (Chapter 13), Colleen Fitzgerald (Chapter 9), and Susan Penfield (Chapter 8). The training of language community members in documentary methods results in a continuous feedback loop between the activities of training, documentation, analysis and revitalization. The benefits of community training are apparent on all levels, from revitalization on a local scale to the recognition of language rights on a global scale (Arka, Chapter 13; Fitzgerald, Chapter 9).

The inclusion of experts from other disciplines in documentation projects has also become a priority in recent years. The value of interdisciplinary documentation should be apparent: including the documentation of music, ethnobiology, and ethnoastronomy, and other domains of knowledge in language projects makes the data richer and more valuable to more stakeholders. Nonetheless, challenges remain, because different disciplines have different goals as well as methods for reaching those goals. Penfield emphasizes that institutions such as universities and journals need to encourage more cross-disciplinary cooperation in order to facilitate the building of interdisciplinary documentation teams which can produce multifaceted data that will be of use to more groups of people.

Ethics has also been a prominent point of discussion over the last two decades (Czaykowska-Higgins, Chapter 11). Importantly, “language documentation is not historically, politically, socially or culturally neutral, and is not simply an intellectual

act (p. 113), and because the work often takes place in small communities that have experienced marginalization, we are obligated to develop ethical professional practices.

A final theme in Part II is the recognition of the role of the newly-elevated status of language documentation in expediting the development of ancillary initiatives to support the field. Most visible among these is the establishment of major funding initiatives specifically targeting documentation and revitalization activities worldwide, as Gary Holton and Mandana Seyfeddinipur (Chapter 10) discuss. These include larger funding regimes like the Endangered Language Documentation Programme (ELDP), the Documentation of Endangered Languages Program (DoBeS), and the U.S. National Science Foundation Documenting Endangered Languages (NSF-DEL) program. As Holton & Seyfeddinipur note, these funding schemes have dramatically improved the quality, scale, and pace of documentation projects around the world.

The development of technological support for language documentation is another area of rapid ancillary progress over the last two decades. Alexandre Arkhipov and Nick Thieberger (Chapter 14) discuss the urgent need to quickly develop software and workflows that adhere to standards for data longevity and interoperability, which has led to an ongoing discussion between linguists, developers, and language speakers/users. In recent years, ease of use has become an additional goal of tool development, which has led to increased participation by members of language communities in documentary data creation, and in the future crowd-sourcing technologies are likely to help ease the transcription bottleneck (Himmelmann, Chapter 4).

The increased awareness of language diversity brought about by language documentation has led to a renewal of efforts to standardize language and dialect names for better cross reference in databases and beyond. Sebastian Drude (Chapter 12) presents the histories of the Ethnologue, Glottolog, the ISO 639 standard, and atlases like the UNESCO Atlas of the World's Languages in Danger. These histories are intertwined and have ramifications for how the general public reacts to declining linguistic diversity worldwide.

4. Part III: Beyond description: Creating and using language documentations

The chapters in Part III present several interrelated themes that center on moving the field forward. In some chapters, this means loosening the reigns of descriptive linguistic research agendas (Riesberg, Chapter 15; Schnell, Chapter 17; McDonnell, Chapter 19), and in others it concerns the broadening of our current understanding of how lexical knowledge is documented (Keren Rice, Chapter 18) and our current conceptions of documentary corpora (Sally Rice, Chapter 16). What ties these chapters together is that they primarily come from an academic perspective and are oriented to outsiders. (This does not mean that they do not recognize the central role of community members, the multipurpose nature of language documentations or the importance of language documentations for language reclamation: chapters by Sally Rice, Keren Rice, and Stefan Schnell all highlight these issues.)

Additionally, the chapters in Part III represent all three elements of the Boasian trilogy, and to different extents they draw attention to how (some part of) each element of the trilogy can be better conceptualized and/or utilized to accomplish desired outcomes in documentary linguistics (i.e., to create a multipurpose record of the practices of a speech community). Bradley McDonnell (Chapter 19), Sonja Riesberg (Chapter 15), and Stefan Schnell (Chapter 17) cover areas that are most likely to fall under the heading of “grammar” (though they also discuss connections between the ‘grammar’ and “texts”; see below), Keren Rice (Chapter 18) covers the “dictionary” through her discussion of

documenting the lexicon, and Sally Rice (Chapter 16) represents ‘texts’ with her discussion of documentary corpora. It is in these detailed treatments of the different components of the Boasian trilogy that we see new (or renewed) manifestations of the connections among each of the elements in contemporary contexts.

Sally Rice’s chapter on documentary corpora is the clearest example of this. She points out that while corpus building and annotation have been widely discussed in documentary linguistics, the characterizations of the applications of a documentary corpus in the language documentation literature—including to a certain extent Himmelmann (1998)—are vague and “replete with elusive and ultimately off-hand comments that do little to clarify exactly what a corpus is capable of” (p. 159). She quite convincingly argues that a fuller understanding of the potential uses of documentary corpora will better motivate the process of creating language documentations. She proposes that “[g]oing forward, we must stop regarding the corpus as a body of recordings, impeccably textualized and identified, and possibly left silent and still in an archive, but instead view it as an active and noisy collection of transcribed conversations teeming with insights about the language and its use that we can eavesdrop on again and again” (p. 161).

Keren Rice isn’t as emphatic in her critique of current practices in the field in her treatment of the lexicon. However, based on Himmelmann’s (1998) discussions of documenting a full array of communicative event types, she points out that while there will be lexical items that cross-cut different event types, there will be other lexical items that are much more likely to occur in one type of event. “Thus, in order to obtain a lexicon that is both broad and deep, a considerable corpus must be developed” (p. 184).

Sonja Riesberg (Chapter 15), Stefan Schnell (Chapter 17), and Bradley McDonnell (Chapter 19) each discuss how current linguistic research—whether that be descriptive grammars, theoretical studies of a single language, or cross-linguistic studies—interfaces with language documentations. How these two activities—documentation and description—interface has been an area of much debate since Himmelmann (1998). These authors, especially Riesberg, list several misconceptions surrounding Himmelmann (1998). McDonnell and Schnell’s chapters additionally provide practical examples of the relationship between linguistic research—including research goals, research questions, and linguistic analysis—on the one hand, and the documentation on which it are based on the other. For example, Schnell argues “it is overall more fruitful for innovative linguistic research to invest into the processing of haphazard language documentation data rather than attempting to collect precisely the kind of data demanded by specific analytical goals” (p. 173).

One final theme that arises repeatedly in this section is the importance of documenting naturally-occurring (connected) speech of different communicative event types. Sally Rice and Bradley McDonnell are most explicit in emphasizing the importance of everyday conversation; Stefan Schnell advocates for the multipurpose functions of documenting traditional narratives, and how such documentation can be (more) useful for linguistic research and (more) desirable for communities. Keren Rice, in her treatment of documenting lexical knowledge, even advocates for a documentary corpus of different communicative event types, as evidenced in her quotation above.

5. Part IV: Fieldwork and language documentation around the world Part IV begins with Claire Bowern’s reflection on linguistic fieldwork (Chapter 20). Her chapter does not summarize the other chapters in this part, nor does it purport to reflect on linguistic fieldwork across the entire world. Rather, from her perspective as one who

conducts linguistic fieldwork in Australia, she reflects on changes in fieldwork practices since Himmelmann (1998) by posing three questions: (i) Have field methods changed? (ii) Has the production of documentation changed? (iii) Has academia changed? She answers each question affirmatively, and in doing so she shows how linguistic fieldwork has progressed in several areas and still needs to develop in other ways. For example, there are more interdisciplinary projects, major increases in the amount of data collected, including conversational and natural data, a greater emphasis on archiving, and better recognition of the critical role of community linguists and language activists, but little has changed in terms of publications and more needs to be done to take advantage of digital media, such as linking media to text. In her conclusion, Bovern points out that while it's difficult to say how many of these changes can be directly attributed to Himmelmann (1998), this influential paper certainly came at the right time in history.

The remaining chapters in Part IV are diverse, both in terms of content and presentation. This reflects not only the different interests and perspectives of the authors, but also the diversity of contexts in which language documentation projects are situated. In many of these chapters, the calls for documentation projects (e.g., by Woodbury 2011, Austin 2014, 2016) to be tailored to their individual contexts are borne out. The most widely discussed issues that show this diversity include:

- (i) collaboration among communities, academic linguists, and other stakeholders, and more generally the role that academic linguists play in documentation projects in different regions, e.g., compare the discussions surrounding collaboration in Ghana (Ameka, Chapter 22), India (Chelliah, Chapter 24), Australia (Singer, Chapter 26), the Pacific (François, Chapter 27), the Chaco (Golluscio & Vidal, Chapter 29), and North America (Rosenblum & Berez-Kroeker, Chapter 32);
- (ii) the degree to which language revitalization/reclamation activities are intertwined with documentations, e.g., compare the differences found in India (Chelliah, Chapter 24), Australia (Singer, Chapter 26), the Pacific (François, Chapter 27), Southern Cone (Zúñiga & Malvestitti, Chapter 28), Mexico and Central America (Pérez Báez, Chapter 31), and North America (Rosenblum & Berez-Kroeker, Chapter 32);
- (iii) the differing needs for training and capacity building in the region, e.g., compare discussions of training needs in India (Chelliah, Chapter 24), eastern Indonesia (Sawaki & Arka, Chapter 25), Amazonia (Beier & Epps, Chapter 30), Mexico and Central America (Pérez Báez, Chapter 31), and North America (Chapter 32);
- (iv) the effects of governmental and other stakeholding institutions in the region and/or official policies on documenting and/or revitalizing Indigenous languages, e.g., compare eastern Indonesia (Sawaki & Arka, Chapter 25), Australia (Singer, Chapter 26), and the Chaco (Golluscio & Vidal, Chapter 29).

Many of these chapters present the history of language documentation and description—and to a lesser extent language reclamation/revitalization—in the region, cataloguing much of what has happened and is currently happening (see, for example, Güldeman (Chapter 21), Ameka (Chapter 22), François (Chapter 27), Zúñiga & Malvestitti (Chapter 28)). Some of these chapters make a particular point in presenting the history of language documentation in the region. Felix Ameka (Chapter 22), for example, presents the history of language documentation and language description in Ghana, making note

of *who* is doing the documentation and/or description. He shows that there has been a shift from outsiders conducting descriptive linguistic fieldwork to insiders and “insider-outsiders” (i.e., those who have knowledge of wider cultural practices and norms of the community, but who are not members of the speech community that is being documented) conducting documentary linguistic fieldwork.

There are also numerous connections between the major themes that surface in the chapters in Part IV and the chapters in earlier parts of the volume. For example, Christine Beier and Patience Epps (Chapter 30) highlight the need for increased training in ethnography as well as contextual training for academic linguists documenting languages in Amazonia, a point that Dobrin & Sicoli (Chapter 5) argue at length. Through their descriptions of individual documentation projects, Golluscio & Vidal (Chapter 29) show how funding initiatives—from DOBES, ELDP, Foundation for Endangered Languages (FEL), and NSF-DEL—have had a major impact on language documentation efforts in the Chaco. At the same time, Sawaki & Arka (Chapter 25) highlight the fact that the vast majority of funded documentation projects in eastern Indonesia are conducted by foreigners, because Indonesian nationals typically do not have the training or capacity to be competitive for such international awards. Both points are discussed at some length by Holton & Seyfeddinipur (Chapter 10) as well.

Finally, some chapters in Part IV take a longer view of language description and language documentation in the region that they cover. For example, Zúñiga & Malvestitti (Chapter 28) discuss the earliest linguistic descriptions during Spanish colonial rule in some languages of the Southern Cone, and Ameka (Chapter 22) discusses early grammars in Ghana, tracing the origins of linguistic fieldwork back to the mid-19th century. Others trace the origins of modern, digital language documentation to activities that were being done prior to the publication of Himmelmann (1998). François (Chapter 27), for example, makes the point that the turn towards documentary linguistics was already taking place in the 1970s “when linguists understood that their role was to record languages in the way they were actually spoken,” which was evidenced by various collections, and in 1996 “Cnrs–LaCiTO created the first online audio archive in endangered languages, ... bringing together valuable fieldwork recordings with their text annotations” (p. 279) (see Jacobson et al. 2001). Circling back to Bowerman’s assessment of Himmelmann (1998), many of the chapters in Part IV show that while much has changed since the appearance of Himmelmann (1998), there was certainly a confluence of activities that led to the development of documentary linguistics as a field in its own right.

6. Some limitations of the volume In concluding this chapter, we ought to mention several limitations of this volume. First, we decided early on to focus on language documentation in this volume, and so there are no chapters dedicated to language reclamation/revitalization. We decided to limit the volume to in this way because we did not want give language reclamation short shrift, and we thought that the topic deserves an entire volume in its own right with editors with more expertise than we have in this area. Interested readers may wish to consult the recently-published *Handbook of Language Revitalization* (Hinton et al. 2018). Note, however, that many of the chapters do discuss issues of language reclamation/revitalization and how multipurpose language documentations relate to such issues. (See in particular Chapter 6 on (de)colonialism.) This to us represents an important trend in documentary linguistics where language reclamation is a driving force behind a language documentation.


Second, in Part IV, we compiled reflections from different regions around the world. One will quickly notice that there are very large gaps. This occurred for two reasons. The first is that we early on realized early on that the volume would become unwieldy if we attempted to provide a representative sample of the world's language. Instead, we aimed to get a diverse set of areas from as diverse a set of authors. The second is that we could not find an author to contribute or the author was not able to contribute a chapter in the end. Thus, our sample of different regions of the world is indeed small and has numerous gaps, but we do hope that these localized portraits of language documentation projects from around the world.

Finally, owing to space limitations we were only able to include vignettes from a small number of the many researchers engaged in documentary linguistics. This fact in itself is reflective of the growing success of the language documentation enterprise. Two decades ago the number of people engaged in documentary work was much smaller, and the views of the discipline—such as they were—could potentially be captured in a single volume. The sheer size of the field today makes this more challenging. Nevertheless, we feel that the views of the contributors to this volume are largely representative and widely shared across the field. No doubt there will be differences of opinion, especially regarding the key issues in language documentation and the future of documentary linguistics as a field. It is our hope that the reflections presented here will help to stimulate future discussion and debate as the discipline of documentary linguistics continues to mature.


References

- Austin, Peter K. 2014. Language documentation in the 21st century. *JournalLIPP* (3). 57–71.
- Austin, Peter K. 2016. Language documentation 20 years on. In Martin Pütz & Luna Filipovic (eds.), *Endangerment of languages across the planet*, 147–170. Amsterdam: John Benjamins.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard Meier, Nicholas Thieberger, Keren Rice & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 57(1). 1–18. doi:10.1515/ling-2017-0032.
- Boas, Franz. 1917. Introductory. *International Journal of American Linguistics* 1(1). 1–7.
- Dobrin, Lise M. & Josh Berson. 2011. Speakers and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 187–211. Cambridge: Cambridge University Press.
- Gawne, Lauren, Barbara Kelly, Andrea Berez & Tyler Heston. 2017. Putting practice into words: Fieldwork methodology in grammatical descriptions. *Language Documentation & Conservation* 11. 157–189.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195.
- Hinton, Leanne, Leena Huss & Gerald Roche (eds.). 2018. *The Routledge handbook of language revitalization*. New York: Routledge.
- Jacobson, Michel, Boyd Michailovsky & John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication* 33(1-2). 79–96.
- Krauss, Michael E. 1992. The world's languages in crisis. *Language* 68(1). 4–10.
- Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.


Bradley McDonnell
mcdonn@hawaii.edu

 orcid.org/0000-0001-6422-2022

Gary Holton
holton@hawaii.edu

 orcid.org/0000-0002-9346-1572

Andrea L. Berez-Kroeker
andrea.berez@hawaii.edu

 orcid.org/0000-0001-8782-515X

**Reimagining Documentary
Linguistics**

Reflections on the scope of language documentation

Jeff Good
University at Buffalo

Language documentation is understood as the creation, annotation, preservation, and dissemination of transparent records of a language. This leads to questions as to what precisely is meant by terms such as *annotation*, *preservation*, and *dissemination*, as well as what patterns of linguistic behavior fall within the scope of the term *language*. Current approaches to language documentation tend to focus on a relatively narrow understanding of a language as a lexicogrammatical code. While this dimension of a language may be the most salient one for linguists, languages are also embedded in larger social structures, and the interaction between these structures and the deployment of lexicogrammatical codes within a community is an important dimension of a language which also merits documentation. Work on language documentation highlights the significance of developing theoretical models that underpin the notion of language, and this can have an impact not only for the practices of documentary linguists but also for the larger field of linguistics. It further suggests that documentary linguistics should not merely be seen as a subfield that is oriented around the collection of data but as one that is in a position to make substantive contributions to linguistic theory.

1. Just what is language documentation?¹ Woodbury (2011: 159) defines *language documentation* as “the creation, annotation, preservation, and dissemination of transparent records of a language.” This definition is undoubtedly useful, and it covers the core goals of most documentary work quite effectively. It also contains within it a set of terms such as *annotation*, *dissemination*, and *transparent* which invite further scrutiny. What level of annotation can be considered adequate? Should dissemination be understood merely in terms of the mechanistic delivery of specific records, or does it require us to think about how records can be used by diverse kinds of users? Who determines

¹I would like to thank two anonymous reviewers for their feedback on an earlier version of this paper.

whether a record is transparent, and who has the burden of ensuring that records are as transparent as possible?

The answers that one may give to questions like these will necessarily play a role in determining the scope of language documentation. Himmelmann (1998) laid out a clear articulation of something that language documentation is not, namely language description, even if these two activities form a natural partnership. In the ensuing decades, documentary linguists have further converged on a set of methods and products that are uncontroversially at the center of language documentation, with the collection of naturalistic recordings of underdescribed languages, accompanied by metadata and annotations consisting of time-aligned transcriptions, translations, and morphological analysis, forming the core of most documentary projects. Indeed, one might view these three components—i.e., recordings, metadata, and annotations—as a “Himmelmannian” trilogy to parallel the Boasian trilogy of grammar, dictionary, and texts.

However, the current stability of this documentary core can lead to a false complacency and to a sense that documentation involves merely repeating the same set of tasks on more and more languages. There is, in particular, a danger that, by deciding in advance that documentation consists of a fixed set of objects, we may fail to notice significant linguistic features of a community that are worthy of documentation but fall outside of what can be captured by the standard approach.

Here, I want to focus specifically on problems that arise from the idea that language documentation involves documenting a “language”, given the ambiguities embodied by this term. The particular concerns that I will raise surrounding just what kind of thing a language is are not new in and of themselves, though my impression is that their implications for documentation are underappreciated. Given that language documentation is ostensibly an activity organized around the idea that there are languages out there in the world to document, it is clear that understanding what we mean by the term *language* has crucial bearing on the scope of the documentary enterprise.

2. What is a “language”?

2.1 Enumeration and language as a set of recorded objects For good reason, the field of linguistics does not operate with a universal definition of *language*. For many kinds of linguistic investigation, the sense of the term is either sufficiently clear from context, or it is not especially relevant. Indeed, Himmelmann (2006: 2) briefly considers this issue with respect to language documentation and argues that a pragmatic approach can be adopted, with work proceeding even in absence of a clear definition.

However, most work within language documentation is directly built on the idea that there is a specific set of languages out there in the world that need to be documented. In a discussion of the rhetoric surrounding endangered languages (which are, of course, the linguistic category that provided the impetus for the development of the contemporary documentary approach), Hill (2002: 127) discusses this in terms of the notion of *enumeration*. This is the assumption that the speech varieties of the world comprise an identifiable set of discrete languages.

This assumption runs immediately into the well-known problem regarding the distinction between languages and dialects. However, where to draw this line in any given case does not raise significant concerns with respect to current approaches to language documentation since the standard techniques are agnostic as to whether the speech variety of focus is classified as a distinct language or not. (The clear exception to this

generalization is the fact that, from the perspective of getting funding to do documentary work, it is much harder to get support to work on an endangered dialect than on an endangered language.)

The enumerative “impulse” can inadvertently lead to the adoption of an assembly-line approach to the task of documentation that is ill suited to local contexts: For each undocumented language, collect a certain number of hours of naturalistic recordings, transcribe and analyze them, make an archival deposit, and consider the language to be “documented” (see also Dobrin et al. (2009) and Austin & Sallabank (2011)). Real-world documentation projects are never so simplistic in their approach. However, highly reductive models are suggested in certain strands of the literature, as seen, for instance, in the description of the Basic Oral Language Documentation method in Bird (2010: 9), which proposes an almost algorithmic approach to collecting data and determining how much annotation is needed. Similarly, Cysouw & Good (2013) develop a definitional scheme that “flips” the usual understanding of the relationship between languages and language resources. Rather than seeing resources as documenting languages that are independently understood to exist, they propose treating collections of resources themselves as defining the language. While the intent of this model is to complement, rather than supplant, more traditional understandings of language, its conceptual foundations clearly rest on a very reductive understanding of what a language is.

2.2 Language as a lexicogrammatical code While the work of language documentation may, at times, lead to an accidental emphasis on the resources produced during the course of documentation over the actual languages themselves, documentary linguists generally operate with a broader conception of language than simply a collection of language resources. However, most work in language documentation still emphasizes a relatively narrow view of language as being constituted by a lexicogrammatical code—that is, as a system of encoding meanings through a combination of lexical elements and grammatical constructions (see, e.g., Woodbury (2011: 177)).

The study of lexicogrammatical codes is at the core of structural approaches to linguistics, and it should hardly be seen as surprising that it has had a central place in work on language documentation. Nevertheless, this approach circumscribes our understanding of what a language is in two crucial ways: First, it ignores the sociolinguistic context in which lexicogrammatical codes operate as a target of documentation (see Childs et al. (2014)). Second, it implies that a language can be defined in terms of a single code rather than as something more complex, such as a set of interacting codes. These points will be developed further below.

The understanding of language as a lexicogrammatical code further implies that there is a potential endpoint to documentation. This is when sufficient data has been collected that the entire code can be revealed through the analysis of the resources that have been collected. This understanding, therefore, represents a conceptual approach where each language is seen as a bounded object, and it, thereby, backgrounds the variation and fluidity that characterize actual language use. This approach to language is analytically powerful and has formed the foundation of modern linguistic analysis since at least the time of Saussure, but it, too, is quite reductive in nature.

There is an additional way in which the lexicogrammatical code approach to documentation is reductive, but this is an incidental aspect of common practice rather than being intrinsic to the conceptual model itself. It tends to result in the privileging of a single code for any given community as being its “true” code. Woodbury (2005,

2011) uses the apt term *ancestral code* to emphasize the fact that most documentary work is nostalgic in orientation, aimed at capturing the properties of some version of a “pure” lexicogrammatical code that has not been impacted by recent patterns contact and language shift, even if such a code never really existed. (See Grinevald (2005) and Dobrin & Berson (2011) for related discussion.)

2.3 Language as a set of interacting lexicogrammatical codes One way in which the equation of a language with a lexicogrammatical code does not align well with real-world patterns of usage involves instances where a set of speech practices that, in some intuitive sense, appear to comprise a language are best understood as being built upon the interaction of multiple lexicogrammatical codes whose opposition to each other is meaningful. A relevant example comes from Kroskrity’s (1992) discussion of Arizona Tewa. In this language, there is a speech register associated with the religious space of the kiva that is highly regulated, with strong constraints on using fixed language. Kroskrity (1992) argues that this pattern of use, a kind of linguistic regulation by convention, is found in different guises in other registers of Arizona Tewa speech, as evidenced, for instance, by prohibitions against code-mixing in everyday speech. While the register associated with the kivas and everyday registers are viewed as elements of the same language and draw on a common lexicogrammatical foundation, their comparison also reveals an important cross-register dynamic of speech regulation. This is manifested in different ways in different registers but appears to be an important feature of the overall linguistic system. Notably, this feature can only be properly documented if one first recognizes the existence of different layers of codes within an overarching lexicogrammatical scheme.

Comparable examples are not hard to find. Storch (2011), for instance, provides extensive discussion of pertinent cases of secret registers found in African languages, and studies of in-law avoidance registers are also relevant, such as the examination of a register of the Nilotic language Datooga known as *gíng’áwêakshòoda*, discussed in Mitchell (2016). This term refers to a speech practice where married women avoid the names of many of their in-laws as well as words that sound like those names. They must replace the relevant words in their own speech, either through the use of conventionalized or semi-conventionalized avoidance vocabulary or other strategies, such as circumlocution. There is one common Datooga grammar among speakers, but the lexicon can differ significantly among them. All speakers must have knowledge of these different lexicons in order to understand each other even if a given individual only uses one of them. This can be modeled as a case where there is a single grammatical code in the language, but multiple lexical ones.

2.4 A lexicogrammatical code with social entailments A more expansive notion of language is at once probably the most usual understanding of the term outside of linguistics and also the one that offers the most complications and opportunities for documentary work: This is the pairing of a lexicogrammatical code (or set of codes, as just discussed above) with social meaning. The range of social meanings that can be assigned to a given language is not an area that appears to be well explored. The most well-known case involves connecting language to culture and nation (see, e.g., Foley (2005: 158)). This linkage is based on an ideology that views language as one manifestation of a deeper ethnocultural essence.

By contrast, Di Carlo & Good (2014) discuss the case of the Lower Fungom region of Northwest Cameroon where a high level of individual multilingualism is found. In that

region, the use of a local language is not understood as linked to essential characteristics of any group, but, rather, primarily serves to index membership in a social group corresponding to one of the local villages. In such a social context, being multilingual allows one to index affiliation to more than one local group, thereby increasing access to resources. In Papua New Guinea, Slotta (2012) discusses the case of the Yopno, who view speech varieties as closely tied to particular locations and as an index of an individual's "sociogeographic" provenance, exemplifying another way that lexicogrammatical codes can be linked to social structures.

These kinds of social entailments connected to the use of a particular lexicogrammatical code can be seen as components of larger language ideologies, and they suggest priorities for documentation within the relevant communities. In Lower Fungom, for instance, the linguistic picture of the region would be incomplete if patterns of multilingualism were not captured. For the Yopno, Slotta's (2012) analysis suggests that instances of language usage where a speaker employs a variety distinct from that associated with their sociogeographic provenance are significant for understanding how social connections are mediated through language.

The methods that dominate language documentation at present are effective at creating records that capture the properties of the world's lexicogrammatical codes. However, they are inadequate for documenting languages if, by this term, we mean not only the codes that comprise a language and their patterns of use but also their social entailments. Capturing the latter requires augmenting the documentary toolkit in ways that can create transparent records not only of lexicogrammatical codes but also of language ideologies, linguistic ecologies, and the sociolinguistic lives of speakers. This would be a challenge, but, as will be further developed below, it is precisely this kind of challenge which demonstrates that language documentation is not merely a check-the-box exercise in data collection but, rather, a proper subdiscipline of linguistics in its own right.

3. Flipping the target: Repertoires rather than languages Language documentation developed within a discipline that treats languages as its primary object of study. Therefore, its focus on languages—however we might define these—is hardly surprising. At the same time, it is also a domain of linguistics that is heavily concerned with speakers (see Grinevald (2007) for one example). Somewhat curiously, though, this concern is not evident in standard approaches to documentary data collection, which tend to view linguistic events, not speakers, as primary (see, e.g., Himmelmann (1998: 168) or the documentary workflow model provided in Thieberger & Berez (2012: 97)). A logical alternative would view the linguistic behavior and knowledge of individuals as the target of documentation. This kind of approach is anticipated in classic works such as Hymes (1962[1971]), which argues for the need for scholarship on the ethnography of speaking (or, as more typically referred to today, the ethnography of communication) to uncover the relationship between the languages of a community and the way the use of those languages patterns in speech, and Gumperz (1964: 137), which develops the notion of *verbal repertoires* understood as "the totality of linguistic forms regularly employed in the course of socially significant interaction."

Documentation taking such ideas as a starting point might, for instance, attempt to make a record of patterns of language usage across time and social setting for a set of speakers associated with a single community rather than emphasizing any particular language of that community. In parts of the world characterized by high degrees of

individual-level and societal multilingualism, such documentation is likely to provide a more accurate record of the linguistic practices of a given speech community than an event-based approach.

This idea is recently considered in detail in the examination of patterns of multilingualism in Africa found in Lüpke & Storch (2013), which points to the possibility of a repertoire-based approach to documentation that can capture the different ways that languages can be known and used in a given community. Lüpke & Storch (2013: 24–27) discuss, for instance, a ritual process intended to improve a woman's chances of successfully having children that involves a significant shift in outward identity. A change in primary linguistic identity is often a part of this ritual

The social meaning of this kind of language shift could never be observed through a purely lexicogrammatical code approach to documentation. Rather, it requires putting the individual's patterns of language use over the lifespan and across different settings in focus. This concern should not be seen as limited to especially salient cases of language shift such as what Lüpke & Storch (2013) describe. Individuals in all speech communities control a range of registers, in some cases actively, in the sense of being able to make use of a given register in their own speech, and, in others passively, in the sense of understanding a given register and knowing its typical range of uses. Some ways of speaking, such as the kinds of linguistic innovations associated with teenagers, may be specifically linked to particular stages of life. Others, such as child-directed speech, may be linked to specific interactional settings. In either case, it is clear that a documentation project which fails to capture these patterns of language in use will result in an impoverished record of a language.

In raising the possibility of a repertoire-based approach to language documentation—that is, one that takes the way individuals use the languages of their communities across time and social spaces as the primary object of study—I do not mean to suggest that this should supersede an event-based approach. Indeed, it would still necessarily require the collection of records of specific linguistic events. However, rather than orienting data collection along the axis of language, it would orient it around the axis of the individual. Pursuing these as two complementary strands of data collection would clearly yield a more transparent picture of the speech practices of a given community than the dominant approach used at present.

4. From language documentation to documentary linguistics The question of the scope of language documentation can, in some sense, be recast as being about the scope of linguistics itself. A complete theory of what it means to create records documenting an entire language will ultimately need to be based on a complete theory of language. Moreover, the fact that language documentation foregrounds the way speakers use language forces it to directly confront issues of the interrelationship between language and culture that many approaches to the study of language set aside. It, thus, leads to an especially expansive view of linguistics.

Terminological fluctuation between *language documentation* and *documentary linguistics* is longstanding, with the two being used apparently interchangeably. The former (and more frequent) term is ambiguous, potentially referring to the activity of documenting a language or the products of that activity. The latter term implies that we are dealing with a genuine subfield of linguistics, requiring theorization, experimentation, and codification in its own right, and that documentary work is not simply a means to some other end, whether this be traditional description, formal analysis, or applied work. The issues raised

here regarding the scope of language documentation, in my view, emphasize the importance of seeing activities surrounding it as belonging to a genuine subfield of linguistics. While the question of just what is a language is of interest to many linguistic subfields, it is clear that language documentation has a special place in answering it. It comes at the question out of a concern for capturing the full range of variation found within the world's lexicogrammatical codes and leads to larger questions of just what it means for a code to be a language at all. Moreover, the discussion here merely scratches the surface of this problem, since little has been said about just what kinds of records are needed to fully and transparently document all the ways that a code can be a language.

Constraints of time, funding, and energy will inevitably cause scholars to model their documentary efforts on the patterns of previous work. While this might allow for the production of good documentary products, it may inadvertently result in a stagnant documentary linguistics. Moreover, it is likely to lead to an impression among the wider community of linguists that documentary linguistics is primarily a "service" subdiscipline, oriented around the collection and dissemination of data to be used for theoretical analysis by specialists in other areas. However, the question of what it means to fully document a language is, ultimately, a complex and theory-driven one. This is a point which documentary linguists should more explicitly acknowledge and convey to the field at large, not only to emphasize that documentary linguistics involves more than mere data collection but also to clarify the kinds of contributions that the subfield can make to theories of language.

I would like to conclude, then, by suggesting that a key challenge for those involved in language documentation is to keep pushing the boundaries of what it means to document the "total linguistic fact" (Silverstein 1985: 220) of a language. Among other things, this would entail not only thinking about the facets of languages that we are already documenting but also those that we are—intentionally or accidentally—omitting from the record.


References

- Austin, Peter K. & Julia Sallabank. 2011. Introduction. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 1–24. Cambridge: Cambridge University Press.
- Bird, Steven. 2010. A scalable method for preserving oral literature from small languages. In Gobinda Chowdhury, Chris Khoo & Jane Hunter (eds.), *The role of digital libraries in a time of global change: 12th International Conference on Asia-Pacific Digital Libraries (ICADL 2010)*, 5–14. Berlin: Springer.
- Childs, G. Tucker, Jeff Good & Alice Mitchell. 2014. Beyond the ancestral code: Towards a model for sociolinguistic language documentation. *Language Documentation & Conservation* 8. 168–191.
- Cysouw, Michael & Jeff Good. 2013. Languoid, doculect, and glossonym: Formalizing the notion ‘language’. *Language Documentation & Conservation* 7. 331–359.
- Di Carlo, Pierpaolo & Jeff Good. 2014. What are we trying to preserve? Diversity, change, and ideology at the edge of the Cameroonian Grassfields. In Peter K. Austin & Julia Sallabank (eds.), *Endangered languages: Beliefs and ideologies in language documentation and revitalization*, 229–262. Oxford: Oxford University Press.
- Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description, volume 6*, 37–52. London: Hans Rausing Endangered Languages Project.
- Dobrin, Lise M. & Josh Berson. 2011. Speakers and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 188–211. Cambridge: Cambridge University Press.
- Foley, William A. 2005. In Peter K. Austin (ed.), *Language documentation and description, vol. 3*, London: Hans Rausing Endangered Languages Project.
- Grinevald, Colette. 2005. Why the Tiger language and not Rama Cay Creole? Language revitalization made harder. In Peter K. Austin (ed.), *Language documentation and description, vol. 3*, 196–224. London: Hans Rausing Endangered Languages Project.
- Grinevald, Colette. 2007. Encounters at the brink: Linguistic fieldwork among speakers of endangered languages. In Osamu Sakiyama Osahito Miyaoka & Michael E. Krauss (eds.), *The vanishing languages of the Pacific Rim*, 35–76. Oxford: Oxford University Press.
- Gumperz, John J. 1964. Linguistic and social interaction in two communities. *American Anthropologist* 66. 137–153.
- Hill, Jane H. 2002. “Expert rhetorics” in advocacy for endangered languages: Who is listening, and what do they hear? *Journal of Linguistic Anthropology* 12. 119–133.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Himmelman, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelman & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.
- Hymes, Dell H. 1962[1971]. The ethnography of speaking. In Thomas Gladwin & William C. Sturtevant (eds.), *Anthropology and human behavior*, 13–53. Washington, DC: The Anthropological Society of Washington.
- Kroskrity, Paul V. 1992. Arizona Tewa kiva speech as a manifestation of linguistic ideology. *Pragmatics* 2. 297–309.

- Lüpke, Friederike & Anne Storch. 2013. *Repertoires and choices in African languages*. Berlin: De Gruyter Mouton.
- Mitchell, Alice. 2016. Words that smell like father-in-law: A linguistic description of the Datooga avoidance register. *Anthropological Linguistics* 57. 195–217.
- Silverstein, Michael. 1985. Language and the culture of gender: At the intersection of structure, usage, and ideology. In Elizabeth Mertz & Richard J. Parmentier (eds.), *Semiotic mediation: Sociocultural and psychological perspectives*, 219–259. Orlando: Academic Press.
- Slota, James. 2012. Dialect, trope, and enregisterment in a Melanesian speech community. *Language & Communication* 32. 1–13.
- Storch, Anne. 2011. *Secret manipulations: Language and context in Africa*. Oxford: Oxford University Press.
- Thieberger, Nicholas & Andrea Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.
- Woodbury, Anthony C. 2005. Ancestral languages and (imagined) creolisation. In Peter K. Austin (ed.), *Language documentation and description*, vol. 3. 252–262. London: Hans Rausing Endangered Languages Project.
- Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.

Jeff Good

jcgood@buffalo.edu

 orcid.org/0000-0001-8679-4654

*“It is simply a feature of a scientific enterprise
 to make one’s primary data accessible to further scrutiny”*
 (Himmelmann 1998: 165)

Reflections on Reproducible Research

Lauren Gawne
La Trobe University

Andrea L. Berez-Kroeker
University of Hawai‘i at Mānoa

Reproducibility in language documentation and description means that the analysis given in descriptive publication is presented in a way that allows the reader to access the data on which the claims are based, to verify the analysis for themselves. Linguists, including Himmelmann, have long pointed to the centrality of documentation data to linguistic description. Over the twenty years since Himmelmann’s 1998 paper we have seen a growth in digital archiving, and the rise of the Open Access movement. Although there is good infrastructure in place to make reproducible research possible, few descriptive publications clearly link to underlying data, and very little documentation data is publicly accessible. We discuss some of the institutional roadblocks to reproducibility, including a lack of support for the development of published primary data. We also look at what work on language documentation and description can learn from the recent replication crisis in psychology.

1. Introduction¹ Himmelmann 1998 seeks to highlight the distinctiveness of language documentation from linguistic description, as well as their “bilateral mutual dependency” (p. 165). Fundamentally, however, the paper is a discussion of the role of data in linguistic analysis. Documentation is the collection and organization of data, and description is the analysis of that data. Himmelmann is adamant throughout the article that the only way documentation and description can be successful is if claims about how a language works

¹Acknowledgments: Our thanks to Peter Austin and Suzy Styles for fruitful discussion about data. Thanks also to our grammar citation and methods collaborators Barbara F. Kelly and Tyler Heston. We would also like to thank our colleagues in the Linguistics Data Interest Group of the Research Data Alliance, particularly our co-chair Helene Andreassen.

can be supported by allowing the reader access to the data on which those claims are founded.

In brief, *reproducibility* (e.g., Buckheit & Donoho 1995; de Leeuw 2001; Donoho 2010) in research means that the data on which publications are based are made available so that other scientists could ostensibly verify the results for themselves. This is distinct from the process of *replication*, in which the steps of a research project are replicated by another scientist, yielding new data which can confirm or contradict previous data. Replication is well-suited to laboratory research, but language documentation data is essentially behavioral data, and analysis of data that is grounded in a specific interactional context that is arguably impossible to replicate. So while replication is not a fruitful aspiration for most documentation-based description, we wholeheartedly agree with Himmelmann that reproducibility is.² Descriptive work needs to be based on sound documentary research methods, and those methods should be made clear by authors of descriptive publications. Relatedly, any published claims about language should be supported by evidence that the audience can, with reasonable considerations for privacy concerns, also access.

Below we examine the context in which Himmelmann (1998) was published, and the developments in linguistics in the last two decades, with regard to the development of digital archiving and open access. We then look at what we can learn as a field from the unfolding crisis of replicability in the field of psychology, and the future of language documentation, description, and data.

Himmelmann (1998) was participating in a larger discussion about the role of data in language documentation, and linguistics more broadly. Sally Thomason, writing as the editor of *Language* in 1994, also articulated concern for clarity regarding the data sources. She called upon linguists to provide “...detailed information about sources of data and methodology of data collection” (Thomason 1994: 413). Some linguists were already actively engaging with data citation in their descriptive linguistic writing. Simpson, at the beginning of her 1983 PhD dissertation on Warlpiri morphology and syntax, states “I have tried to indicate the source of each example sentence where I know it. If the example sentence is made up, I have indicated this, unless the sentence is elementary” (1983: 4). Data citation is an important feature of reproducible research, but it is only of use if the interested reader can resolve that citation to the original data. Digital archives have provided an important development in data sharing.

2. Development of archives One of the most immediately obvious developments in documentation since Himmelmann (1998) is the network of digital archives that provide a persistent and secure location for the storage of linguistic data. Himmelmann voices his concern that “In recent decades, hardly any comprehensive collections of primary data have been published” (1998:164), a concern that is objectively no longer true thanks to the rise of digital archiving. The permanent preservation of one’s materials, once a privilege reserved for only the most senior linguists, is now a common part of the documentary linguist’s workflow.

While analog language archiving had been part of anthropological practice since the late 19th century, the development of digital archiving methods for language documentation began in earnest in the early 21st century (see e.g., Woodbury 2011; Henke & Berez-Kroeker 2016). Those years saw the rise of funding schemes for language

²See Berez-Kroeker et al. 2018 for a discussion on reproducibility in linguistics in general.

documentation like DoBeS³ in 2000 and ELDP⁴ in 2003, both of which provided a repository for preserving their grantees' work. The NSF-funded Electronic Metastructure for Endangered Languages Data (EMELD)⁵ project provided much-needed education to linguists on how to digitally preserve language documentation (Boynton et al. 2010).

Alongside archiving has come a standardization of metadata. Himmelman does not actually use the term 'metadata'—instead the article refers to “information to be included” (1998: 189, see also 169–170)—while discussing what has now become known as 'metadata' at some length. In the early 2000s, the Open Language Archives Community (OLAC)⁶ was building a metadata standard based on Dublin Core specifically for describing digital language materials (e.g. Bird & Simons 2003). Similarly, the International Standards for Language Engineering Metadata Initiative (IMDI)⁷ was developed in the DoBeS context.

Digital archives not only provide persistent data storage, but they also provide access to the data thanks to improvements in the internet. Himmelman is sensitive to placing speaker attitudes at the centre of archiving models with a focus on controlled access (1998: 171–175, 189). There has been considerable discussion about ethics and access to documentation materials (Dwyer 2006; Garrett & Conathan 2009; Macri & Sarmento 2010; Shepard 2016), and some archives have implemented different levels of accessibility to materials (eg Green et al. 2011; Nathan 2010; 2014). This is an ongoing conversation, as internet access is still not globally balanced, and speakers of many of the languages represented in archives are unable to view or use deposited materials through lack of access. In terms of reproducible research at least, we have solved many of the barriers that were a concern in 1998, and can now do a lot more than meet Himmelman's minimal solution of providing an “edited version of the fieldnotes” (1998: 165).

3. Open Access The Open Access movement (OA) was beginning to coalesce in the late 1990s, and has been an important influence on the development of archiving practice in documentation. In 1997 the Association of Research Libraries developed the Scholarly Publishing and Academic Resources Coalition (SPARC),⁸ which had an early focus on encouraging open access journal publishing.⁹ OA radically altered the publishing landscape (Joseph 2013), and we see that effect today, with journals like *Language Documentation & Conservation*¹⁰ and presses like *Language Science Press*¹¹ that cost nothing to authors or readers. The OA movement is now actively involved in encouraging open access data practices (SPARC n.d.; Kitchin 2014). Language documentation archives have been leaders within the humanities and social sciences when it comes to advocating for open access, or at least mixed access for different uses.

OA publication has been aided by the creation of Creative Commons (CC) licenses.¹² Founded in 2001,¹³ CC allows copyright holders to specify how members of the public

³<http://dobes.mpi.nl/>

⁴<http://www.eldp.net/>

⁵<http://emeld.org/>

⁶<http://www.language-archives.org/>

⁷<http://tla.mpi.nl/imdi-metadata/>

⁸<http://sparcopen.org/>

⁹<http://sparcopen.org/our-work/research-data-sharing-policy-initiative/>

¹⁰<http://nflrc.hawaii.edu/ldc/>

¹¹<http://langsci-press.org/>

¹²<http://creativecommons.org>

¹³<http://creativecommons.org/about/history/>

may and may not use their work, including whether attribution is required and whether commercialization is allowable. The licenses are both machine- and human-readable for ease of use. Many archives now use CC licenses for OA data. The CC framework provides a scaffold for discussions between language documentation researchers and communities, making this issue somewhat easier to navigate than it was when Himmelmann was writing (1998: 175).

4. Data sharing in today's practice While archives have provided a robust way to share data, we still are not seeing complete uptake of archiving or other practices that lead to reproducibility. In a survey of one hundred descriptive grammars published between 2003 and 2012 that we conducted with Barbara F. Kelly and Tyler Heston, we found that data archiving before publication was only mentioned in 22 publications, and only eight publications included data citation that resolved back to a locatable corpus (Gawne et al. 2017). Many published grammars in our survey do not discuss basic methodological information like the number of speakers who contributed, or recording equipment used, which prevents the reader from understanding the nature of the data on which analysis is built. In a similar vein, Thieberger (2017) looked at 1,708 grammars published since 1967 and found that for 1,253 of the languages there were fewer than 40 items in an OLAC archive, indicating that for the vast majority of descriptive grammars the primary data on which they are based cannot be found or used.

Language documentation has become a field with its own journals, conferences, network of archives and funding, but there remains a fundamental disconnect between documentation data and subsequent description. A major reason for this is the fact that the academic environment does not provide incentives for good practice in reproducibility. We add our voices to Himmelmann's in seeking better transparency in research methodology to ensure that readers can better judge the "reliability, naturalness, and representativeness of the data" (1998: 162), and we believe the best way to do this is through archiving and citation.

Preparing data for archiving is a time-consuming process that is not viewed as having academic merit on par with published analyses. Management and curation of data for archiving is a time-consuming process, even when the documentation workflow is set up to optimize the process. This means that even the best-intentioned documentation practitioner can find themselves with a large amount of work to do that is undervalued by university hiring, tenure and promotion committees. Descriptive work, in contrast, results in peer-reviewed publications, which are still the primary yardstick for measuring academic productivity.

The status quo is changing to some extent. Some initiatives have sought to use the current incentive structure to give recognition to documentation work, such as *Dictionaria*, which uses a peer-reviewed model for digital dictionary databases;¹⁴ the *Language Contexts* series in *Language Documentation and Description*, which publishes contextualising metadata for a language;¹⁵ and the publication of descriptions of archival collections in *Language Documentation & Conservation*, which act as citable proxies for datasets within current citation mechanisms (e.g. Salfner 2015).

Other efforts have been directed at raising the profile of documentation and corpus building. In 2010 the Linguistic Society of America passed the *Resolution Recognizing*

¹⁴<http://home.uni-leipzig.de/dictionaryjournal/about-the-journal/>

¹⁵www.e-publishing.org/language-contexts

the Scholarly Merit of Language Documentation, which recognized corpora and other documentation outputs as “scholarly contributions to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty.”¹⁶ In a similar spirit, the DELAMAN Franz Boas Award “recognizes and honours junior scholars who have done outstanding documentary work in creating a rich multimedia documentary collection of a particular language that is endangered or no longer spoken.”¹⁷ While it is important that there are positive motivators for archiving, a great deal of the archiving undertaken in recent years stems from a more prosaic motivation: funders increasingly require data to be archived, and open access where feasible, as part of the funding process (Austin 2014).

We are still grappling with the question of how to assess the quality of archival collections. While we acknowledge the existing peer review mechanisms for publications are not without their failings, as a discipline we have not yet come up with a commonly agreed-upon way to assess the quality of documentation collections (though note the recent draft for a *Statement on the Evaluation of Language Documentation for Hiring, Tenure, and Promotion*¹⁸ by the Linguistic Society of America and work of the committee of the Australian Linguistic Society, reported in Thieberger et al. 2016). Himmelmann also observed the need to assess documentation work (1998: 181). He focuses mainly on different types of data collection, such as elicitation, tasks and different genres of spontaneous text (1998: §3.3), however there are many factors that need to be considered including quality of recordings, number of speakers, presence of video data, and quality of metadata (Woodbury 2014; Thieberger et al. 2016).

5. Data citation in today’s practice While the move towards accessible corpora has been one challenge, another has been the lack of citation of that documentation data in publications. Editors and publishers, for the most part, have not made explicit an expectation to cite examples of linguistic phenomena (sentences, lexical items, etc.) back to the dataset whence they came. While most linguists would never dream of quoting from another author’s work without a proper citation, those same scholars will happily quote from their own extensive corpora without any citation whatsoever.

We believe in the need for data citation, and have been working alongside our colleagues in the Linguistics Data Interest Group of the Research Data Alliance,¹⁹ to bring these beliefs together in a document known as the *Austin Principles of Data Citation in Linguistics*.²⁰ At the core of these principles is the belief that “[l]inguists should cite the data upon which scholarly claims are based.” (Berez-Kroeker et al. 2017), a belief that echoes the quote from Himmelmann in the epigraph to this chapter. Data citation can help the researcher return to the original data to confirm hypotheses as analysis develops, and it can also help a reader locate the example in the corpus, to seek more contextual information to reproduce the original hypothesis, or for an analysis that the original data was not necessarily presented with a focus on (e.g. looking at the case-marking in a sentence that was originally used to exemplify a feature of tense). Although we are accustomed to seeing example sentences presented as written artefacts, we agree

¹⁶www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

¹⁷<http://www.delaman.org/delaman-franz-boas-award/>

¹⁸www.linguisticsociety.org/content/draft-lsa-statement-evaluation-language-documentation-hiring-tenure-and-promotion

¹⁹<http://rd-alliance.org/groups/linguistics-data-ig>

²⁰<http://site.uit.no/linguisticsdatacitation/>

with Himmelmann that most contextualization of the utterance is lost in print, including prosody and gesture, as well as the possibility to “gloss over” complexities (1998: 191 fn5).

Citing your own data encourages others to cite your data in their work as well. Himmelmann notes that any language documentation corpus includes information well beyond the scope of what a single researcher or team can undertake to analyze (1998: 163). We have seen very little uptake of documentary data in descriptive work published by researchers other than the data collector(s); in our study of articles published in *Linguistic Typology* between 2012–2017 we found that the overwhelming majority of authors draw on published descriptions or their own documentation data (Gawne et al. 2017), but almost never the datasets of other data collectors. We believe that ultimately the citation of data will become standard practice, through editorial policies that make it a norm like other forms of citation.

6. The replication crisis in psychology Linguistics is not the only field to have considered the role of data and analysis in research. The ‘replication crisis’ that started in medical science (see Goldacre 2010 for a summary) and is now being played out in social psychology (Chambers 2017) has much to teach us about the importance of transparency in research methods and data presentation, as well as how we can best approach these themes as a community of researchers. As we discussed above, we do not believe that language documentation and description should strive for replication, which is more relevant to psychology, but there are lessons in this crisis for the future of reproducibility as well. Psychology, like linguistics, is interested in thresholds, not absolutes, in the often-difficult to establish nature of human behaviour.

In 2011 Daryl Bem, a social psychologist, published a paper that demonstrated, across a series of experiments, statistically significant effects of ‘precognition’, with participants appearing to contradict the flow of time and show priming effects on early parts of the experiment based on later parts of the experiment (Bem 2011). The research methods were all meticulously reported, leaving the reviewers to either decide that ‘precognition’ did exist, or the methods of social psychology were not reliable. Bem’s work appears to have been shaped by a ‘forking paths’ analysis (also known as ‘experimenter degrees of freedom’), where each decision in the analysis process appears to be sensible, in keeping with the norms and best practice of the field, but helps the researcher converge on the outcome they want. Bias in each step of data collection can lead to bias in the analysis, which can lead to bias in the meta-analysis that shapes the trajectory of the field. In linguistics, we’ve seen that some topics have been neglected as specific targets for documentation work, because they’ve been considered marginal to a particular conceptualization of language. These phenomena may eventually be shown to be less marginal than had been originally thought (e.g. ideophones, see Dingemanse et al. 2018).

Bem’s case is egregious because believing his findings contradicts the basis of causality on which our understanding of the universe is built, but there are a number of other questionable research practices that the field of psychology is critically analysing. One of these is *hypothesising after results are known*, or HARKing—where the narrative for the data is often changed to fit a more compelling hypothesis after collection is complete and analysis has begun (e.g. deciding that the variable of gender is the significant difference, even though that wasn’t the original aim of the experiment). The other is *p-hacking*, running numerous statistical processes to ‘find’ results in the data, which then leads to HARKing to create a publishable narrative (see, for example, the ‘pizzagate’ controversy surrounding work by Brian Wansink and colleagues (problems with this

research summarized in van der Zee et al. 2017), in which data about pizza consumption was sliced (like an unethical pizza) into statistically-significant subsets to fit the research narrative). Although most descriptive work does not require the explicit formation of hypotheses, researchers do have a set of expectations about what linguistic features a language might demonstrate, based on the typological profiles of related languages. Similarly, a researcher must always select the example sentences to illustrate a descriptive grammar, which is by no means an objective process. Providing the reader with additional examples through presentation of the original data can help mitigate these limitations of descriptive work, in the way that pre-registering hypotheses and presenting data sets is helping in the field of psychology.

The crisis of replicable methodology in psychology was the motivation for Brian Nosek and hundreds of colleagues to attempt to replicate 100 experimental psychology studies published in 2008 (Open Science Collaboration 2015). Fewer than fifty percent of the studies were successfully replicated. In 2013, while working on the replication study, Nosek and colleagues started the Center for Open Science,²¹ a researcher-driven organization that builds easy-to-use tools and protocols, as well as leading discussions about the nature of research practice.

7. Looking ahead The problems in psychology arose in part because research practices were not transparent enough. Researchers generally were not required to present their methodology in a way that ensured replication, nor to commit to a course of research, maintain it through to publication and share the underlying data. In recent years, the move towards greater transparency in psychology has included ‘pre-registration’ of methods, either as a peer-reviewed process that becomes the first half of the final peer-reviewed paper,²² or as a non-reviewed methodology that is time-stamped and limits the ‘researcher degrees of freedom’ that can influence the final outcomes.²³

Language documentation does not have the same experimental focus, so we would not want pre-registration as a solution to our research problems, nonetheless we still generally don’t make it easy for our readers to access the datasets on which our analyses are based and are therefore equally susceptible to research pitfalls. Even Thomason noted during her tenure as editor of *Language* that erroneous data “occur[ed] frequently—so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable” (1994: 409). We need to continue to develop a social and technological infrastructure in linguistics that allows us to reap tangible rewards for the creation, management, and citation of linguistic data as much as we do for linguistics publications. In short, we still don’t value language *documentation* as much as we value linguistic *description*. We can do better.

Language documentation has changed over the last 20 years thanks to the development of digital data collection methods, and online archives that allow for both the storage and dissemination of recorded materials. While we have made some moves towards a more open approach to data that would support research reproducibility, there is still more work to be done to ensure that the link is made clear between linguistic description and the documentation that it is based upon. At a minimum, we believe

²¹<http://cos.io/>

²²Very recently Timo Roettger of Northwestern University has put together an initiative to encourage more linguistics journals to adopt Registered Reports (RRs). For information on the initiative see <http://linguistlist.org/issues/29/29-3168.html>

²³<http://cos.io/prereg/>

that all data from documentation should be archived with a digital repository that has a mandate for long-term storage. Where the data are not sensitive or controversial, they should be made accessible to both the language speakers and to researchers who wish to confirm existing analyses, test new analyses or explore previously under-described phenomena in the language. Descriptive work should clearly state the research methods used in collecting the data that forms the basis of the research, make clear where the data are located and should explicitly link each piece of data to its place in the documentation data. Digital archives and the Open Access movement have given us the tools to make this happen. When all of this is common practice, and not just the practice of a subset of researchers, we will have made a clear move in the direction of reproducible research.

References


- Austin, Peter K. 2014. Language documentation in the 21st century. *JournaLIPP* 3. 57–71.
- Bem, Daryl J. 2011. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* 100(3). 407–425. (doi:10.1037/a0021524)
- 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. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. (doi:10.1515/ling-2017-0032)
- Bird, Steven, & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 57–582.
- Boynton, Jessica, Steven Moran, Helen Aristar-Dry & Anthony Aristar. 2010. Using the EMELD School of Best Practices to create lasting digital documentation. In Lenore A. Grenoble & Louanna Furbee-Losee (eds.), *Language documentation: Practice and values*, 133–146. Amsterdam, Philadelphia: Benjamins.
- Buckheit, Jonathan B. & David L. Donoho. 1995. WaveLab and reproducible research. In Anestis Antoniadis & Georges Oppenheim (eds.), *Wavelets and statistics*, 55–81. New York: Springer.
- Chambers, Chris. 2017. *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton: Princeton University Press.
- Dingemanse, Mark. 2018. Redrawing the margins of language: Lessons from research on ideophones. *Glossa* 3(1). 1–30. (doi:10.5334/gjgl.444)
- Donoho, David L. 2010. An invitation to reproducible computational research. *Biostatistics* 11. 385–388.
- Dwyer, Arianne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann, & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Mouton de Gruyter.
- Garrett, Andrew & Lisa Conathan. 2009. Archives, communities, and linguists: Negotiating access to language documentation. Presentation at the *Linguistic Society of America Annual Meeting*. (http://www.ailla.utexas.org/site/lisa_olac09/conathangarrett_lsa_olac09.pdf)
- Gawne, Lauren, Andrea L. Berez-Kroeker & Helene N. Andreassen. 2017. Data citation in linguistic typology: Working towards a data citation standard in linguistics. Presentation at *Association for Linguistic Typology 12*. Canberra: December 11–15.
- 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.
- Goldacre, Ben. 2010. *Bad science: Quacks, hacks, and big pharma flacks*. London: McClelland, Stewart.
- Green, Jennifer, Gail Woods & Ben Foley. 2011. Looking at language: Appropriate design for sign resources in remote Australian Indigenous communities. In Nick Thieberger, Linda Barwick, Rosey Billington & Jill Vaughan (eds.), *Sustainable data from digital research: Humanities perspective on digital research*, 66–89. Melbourne: Custom Book Centre, The University of Melbourne.

- Henke, Ryan E. & Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation & Conservation* 10. 411–457.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 6. 161–195.
- Joseph, Heather. 2013. The Open Access Movement Grows Up: Taking Stock of a Revolution. *PLoS Biol* 11(10). (e1001686.doi:10.1371/journal.pbio.1001686)
- Kitchin, Rob. 2014. *The data revolution*. London: Sage.
- Leeuw, Jan de. 2001. Reproducible research: The bottom line. *UCLA Department of Statistics papers*. (<http://escholarship.org/uc/item/9050x4r4>) (Accessed 28 March 2018)
- Macri, Martha & James Sarmiento. 2010. Respecting privacy: Ethical and pragmatic considerations. *Language & Communication* 30(3). 192–197.
- Nathan, David. 2010. Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing* 4(1–2). 111–124.
- Nathan, David. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. *Language Documentation and Description* 12. 187–208.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716. (doi:10.1126/science.aac4716)
- Salfner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267.
- Shepard, Michael Alvarez. 2016. The value-added language archive: Increasing cultural compatibility for Native American communities. *Language Documentation & Conservation* 10. 458–479.
- Simpson, Jane. H. 1983. Aspects of Warlpiri morphology and syntax. Massachusetts Institute of Technology PhD dissertation. (<http://hdl.handle.net/1721.1/15468>) (Accessed 2018-03-18)
- SPARC. n.d. Open Data Factsheet (11.10-2). (<http://sparcopen.org/open-data/>) (Accessed 2018-04-2012)
- Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36. 1–21. (doi:10.1080/07268602.2016.1109428)
- Thieberger, Nick. 2017. LD&C possibilities for the next decade. *Language Documentation & Conservation* 11. 1–4.
- Thomason, Sarah. 1994. The editor's department. *Language* 70. 409–423.
- van der Zee, Tim, Jordan Anaya & Nicholas J. L. Brown. 2017. Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. *PeerJ Preprints* 5:e2748v1. (doi:10.7287/peerj.preprints.2748v1)
- Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–211. Cambridge: Cambridge University Press.

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


Lauren Gawne

l.gawne@latrobe.edu.au

 orcid.org/0000-0003-4930-4673

Andrea L. Berez-Kroeker

andrea.berez@hawaii.edu

 orcid.org/0000-0001-8782-515X

Meeting the transcription challenge

Nikolaus P. Himmelmann
Universität zu Köln

The major challenge for language documentation in the next decade or two is what could be called the *transcription challenge*. This is a multilayered challenge that goes far beyond the practical challenge of speeding up the transcription process. Transcription, as practiced in language documentation, involves language making and changes the language ecology. Despite its centrality to language documentation, transcription remains critically undertheorized and understudied. Further progress in language documentation, and ultimately also its overall success, crucially depends on further investigating and understanding the transcription process, broadly conceived.

1. Phonetic transcription, discourse transcription and the transcription bottleneck¹ In linguistics, the term *transcription* is perhaps most closely associated with (narrow) **phonetic transcription**. The *International Phonetic Alphabet* (IPA) provides a means to capture core characteristics of the articulatory movements involved in speaking and thus to represent spoken language in writing. The transcription challenge discussed here does not pertain to this practice, as most challenges for phonetic transcription have successfully been resolved over the last century and a half, at least with regard to the segmental level. As a background to what follows, however, it would do well to remember that developing the IPA required solving major conceptual and practical issues. Perhaps more importantly, it would do well to remember that the representation of the suprasegmental aspects of speech continues to be a challenge. See Ladd (2014) for pertinent discussion.

Another level of transcription pertains to the creation of a written representation of recordings of more or less natural communicative events (everyday conversations, narratives, interactive games, speeches, etc.). The written representation typically provides the basis for the further analysis of such events. Transcription practices in this domain have been the object of theoretical and practical reflections in (different traditions

¹I am very grateful to two anonymous reviewers and editor Bradley McDonnell. In my experience, it has been rare to get such perceptive and helpful input in the reviewing process. Thank you very much! Thank you very much also to Katherine Walker for thoroughly editing English grammar and style.

of) discourse and conversation analysis, Elinor Ochs' (1979) paper *Transcription as Theory* being the classic example (cp. Edwards & Lampert 1993 for a collection of papers on this topic and Bucholtz 2007 for a more extensive bibliography). Unfortunately, neither the theoretical concerns nor the practical guidelines developed in these traditions (e.g. Du Bois et al. 1992, Selting et al. 2009) have had a major impact on practices in field linguistics and language documentation. That is, despite the fact that **discourse transcription** is at the core of documentary linguistic activity, it remains a topic that is rarely discussed in the field.² Consequently, there is little agreement about very fundamental decisions such as how to segment spoken language (cp. Himmelmann 2006 for a short overview of the main issues). More often than not, segmentation units above the word (i.e. prosodic units and/or syntactic phrases) are not explicitly discussed or justified, and are thus difficult to reconstruct and evaluate for users of a documentation. To make discourse transcription a major topic in the field, then, is one aspect of meeting the transcription challenge in language documentation.

A further, related aspect of this challenge is the **transcription bottleneck** (cp. Seifart et al. 2018). Given state-of-the-art recording technologies and a community supportive of creating a comprehensive record of their speaking practices, it is now relatively easy not only to compile a largish collection of documentary recordings, but also to archive them and make them available to other interested parties. But making them truly accessible by adding transcription and translation is a different matter altogether. Estimates of the factor involved here vary, depending on recording quality, the number of speakers involved, etc. Factors smaller than 10 (i.e. ten minutes are necessary to transcribe and translate one minute of recording) are rarely mentioned, and factors as high as 150 and higher are not unrealistic in the case of complex multiparty conversations. Some aspects of the transcription-cum-translation process are fairly mechanical and highly repetitive. Without doubt, support from machine-based speech processing could be of great help in speeding up the transcription process.

A number of efforts in this regard have been undertaken in recent years, as of yet without major success.³ Given the relatively small amounts of data and manpower typically available in the case of underdocumented languages, it is clear that any success of such efforts will not even remotely approach the power of the automatic transcription tools for natural speech currently emerging for major national languages, in particular English. But even the automatization of tasks such as identifying different speakers (speaker diarization), proposing an initial rough, pause-based segmentation and recognizing high-frequency items and phrases would already be of major help in processing documentary recordings. An important side effect of these efforts, inasmuch as they are not confined to mere phone recognition, is the fact that they force the community to become more explicit about its transcription practices, thus addressing the issue of discourse transcription mentioned above.

²To my knowledge, Crowley (2007:137-141) is the only work on linguistic fieldwork that discusses the practicalities of transcribing larger amounts of narrative and conversational speech, going beyond the problems of basic procedure and properly capturing sound. In other subfields of linguistics, (introductory) discussions of transcription tend to be considerably more comprehensive and sophisticated. See Nagy & Sharma (2013) for an example.

³Including the *Transcription Acceleration Project* at the Australian Centre of Excellence for the Dynamics of Language (<http://www.dynamicsoflanguage.edu.au/news-and-media/latest-headlines/article/?id=early-results-from-survey-exploring-transcription-processes>) and the *Kölner Zentrum Analyse und Archivierung von Audiovisuellen Daten* (<http://ifl.phil-fak.uni-koeln.de/32830.html>).

2. The real challenge: understanding the transcription process However, the points mentioned so far are only the beginnings of what I consider to be the core of the transcription challenge: reaching a **better understanding of the transcription process** itself and its relevance for linguistic theory. The two central questions here are:

- (1) What do speakers and researchers actually do when they transcribe (or assist in transcribing)?
- (2) How does the fact that transcription converts specimens of spoken language into a written representation—possibly accompanied by further annotations such as a translation and notes on grammar and cultural background—affect the overall language ecology in the community providing the specimens of natural speech compiled in a language documentation?

As to the first question, it would be rather naïve to consider transcription exclusively, or even primarily, a process of mechanically converting a dynamic acoustic signal into a static graphic/visual one. Transcription involves interpretation and hence considerable enrichment of the acoustic signal. That is, transcription necessarily involves hypotheses as to the meaning of the segment being transcribed and the linguistic forms being used. Linguistic forms tend to be underdetermined by the acoustic signal, as everyone who has ever engaged in transcribing spontaneous speech knows (missing or unclear segments, ambiguous reduced forms, etc.; see also Hermes & Engman (2017: 65 *passim*)). But how exactly does interpretation and enrichment actually work in the transcription process?

In Jung & Himmelmann (2011), we provide some preliminary observations concerning the transcription process, based on recorded transcription sessions where a linguist works together with a native speaker. Typical reactions to the transcription task, bearing witness to the creative aspects of the transcription process, include the tendency for the native speaker to paraphrase what is said rather than repeating it more or less directly. There is also the tendency to edit out elements typical of spoken language such as particles, hesitations and the like. The converse tendency is to edit in material that is deemed to make the transcript ‘better’, ‘more correct’ or ‘clearer’, such as using fuller verb forms, pronouns, and so on. Similar changes occur when transcripts are further edited for publication (Mosel 2014). Marten & Petzell (2016) give a very instructive example of the kinds of ‘purifications’ that often occur in multilingual settings with a major dominant language (in their case study Swahili).

Furthermore, in most instances, transcription also involves language learning. In fact, as illustrated and further discussed in Hermes & Engman (2017), transcription may be used as one way to learn a language, especially when younger speakers collaborate with older speakers in a transcription task. As a consequence, there are overlaps between the kinds of processes that typically occur in (adult) second language acquisition and transcription, and hence there is major potential for cross-fertilization between these two fields of linguistic inquiry.

In order to better understand the transcription process and thus to supply an answer to question 1 above, these and other aspects of the transcription process need to be studied much more systematically. Given the fact that written transcripts are underdetermined by the acoustic signal, transcriptions of the same recording produced by different transcribers/transcription teams will differ in some details. Consequently, questions such as the following arise: How variable are the transcriptions typically produced in language documentation, i.e. to what extent are they underdetermined by the acoustic signal (which of course itself will vary across recordings)? What is the potential impact of this variability for subsequent analyses? What kinds of phenomena are particularly salient for transcribers and how do these intersect with the phenomena salient for second language learners? In order to answer these questions, we need transcription experiments that target those features of transcripts that are particularly prone to variation and those that tend to prompt special attention. Bucholtz (2007), for example, investigates variation in format choices, orthographic variation and variation in translation. Himmelmann et al. (2018) is another example that investigates a feature well-known to be highly variable, namely prosodic segmentation.

However, while better understanding sources of variability and variation in transcription is important for putting language documentation on a safer methodological footing, something else may potentially be of even greater significance, not only for language documentation but for the language and cognitive sciences more generally. When speakers edit in and edit out, what kind of knowledge and norms do they base their decisions on? Why is it that all over the world, speakers of very different languages, living in very diverse linguistic settings, have clear ideas about the fact that some parts of a recorded spontaneous utterance are ‘not relevant’ and hence should be edited out? And that the acoustically observable form X is the short/reduced variant of form Y? And that expression A is ‘better’ and ‘more complete’ than expression B?

The first, seemingly trivial, answer that springs to mind is that all of these ideas and reactions are based on a written standard and prescriptive traditions learned in schools. But then, what about linguistic varieties without a written standard and not used in (formal) education? Of course, very few, if any, settings exist nowadays where there are not at least a few speakers of a given variety who are also familiar with a written language standard and have received some formal schooling. Hence, it is possible that ideas as to how written language should look and what the proper ways of speaking are have disseminated from written language traditions and schooling. But does this really suffice to explain that speakers who are illiterate and do not know a regional or national standard(ized) language show the same tendencies for editing in and out when assisting with transcription?

An alternative explanation for the typical reactions shown by speakers when assisting in a transcription task is that transcription taps into a form of linguistic knowledge that differs from the linguistic knowledge underlying linguistic behavior in spontaneous interactions. Transcription is of course not part of anyone’s native linguistic repertoire—it is a new way of dealing with language for those who engage in it for the first time. However, while being new in its specifics, it probably belongs to a larger class of activities that involve **metalinguistic knowledge and awareness**, i.e. a reflective mode (as opposed to a production mode) in dealing with language. Other examples where this mode comes into play are language games, verbal arts, adult language acquisition and conscious choices made in multilingual settings. If metalinguistic knowledge and awareness are factors in transcription, then, obviously, investigations of the transcription process should

be informed by research on metalinguistic knowledge and awareness in other domains.⁴ At the same time, transcription opens a new venue for researching this type of linguistic knowledge.

3. Productive and reflective modes of language use It is well known that there are fundamental differences between speaking and writing, with regard to both the linguistic structures being used and the cognitive resources that are deployed in the respective production processes (e.g. Chafe 1982, 1994; Akinnaso 1982, 1985; Biber 1988, 2014). The hypothesis proposed here is that these differences are not exclusively bound to the differences in the communicative channel (auditory vs. visual) but also occur in the oral language domain itself, as also assumed in Labov's (1972) notion *attention to speech* and in much contemporary work on 'style' (e.g., Coupland 2007).⁵ That is, there are language-related activities, including language games, verbal arts and transcription, that make use of a reflective mode in dealing with (spoken) language, which is different from the mode employed in producing spontaneous speech. This hypothesis and the considerable body of work on the differences between speaking and writing, then, provide the basic framework for investigating what speakers actually do when transcribing, and what this implies for linguistic and, more generally, cognitive theories.

A demanding, but also exciting, research program follows when transcription is approached in this way. Transcription practices need to be documented more systematically in order to get the full picture of which strategies and which forms of knowledge are applied in transcription.⁶ Other linguistic activities which tap into metalinguistic knowledge and awareness need to be identified to allow us to develop a comprehensive view of the reflective use of language. As metalinguistic knowledge and awareness are not accessible to direct observation, many methodological challenges have to be overcome in determining how insights about this knowledge type can be derived from observable linguistic behavior (such as the behavior shown by native speakers in transcription tasks) and tested in well-designed experiments.

4. Transcription as language making Inasmuch as transcription necessitates the development of a new way of representing speech, it is a prototypical instance of language making.⁷ This aspect of transcription is often underestimated, because many practitioners tend to underestimate, and hence fail to take into account, the amount of interpretation and enrichment involved in the transcription process as highlighted in the preceding two sections. But, in addition to the cognitive aspects discussed above, when its language-making potential is taken seriously there is also a social aspect to transcription.

⁴The journal *Language Awareness* provides examples, which, however, mostly concern the North American and European settings and frequently pertain to the role of awareness in language learning and teaching. See Verschik (2015) for a brief survey of the role of language awareness in multilingual settings and language contact.

⁵There is an almost forgotten tradition dating back to the 19th century that has developed this hypothesis, using the term *Ausbau* for what here is called *reflective mode*. Maas (2009, 2010) provides details and arguments.

⁶In fact, as one reviewer rightly points out, current practice is often wanting even on the most basic level of including the information of who was involved in the transcription (only a member of the community? the researcher? both?).

⁷Note that even in communities with well-established writing traditions both spontaneous and scientific representations of spoken language usually involve the invention of new conventions such as new uses of punctuation and various attempts to capture salient aspects of speech by indicating lengthening, stress and frequent fast-speech forms (e.g., *gonna*), etc.

Transcription introduces a new element into the linguistic repertoire of those who engage with it. What is completely unclear to date is whether and to what extent this change may have repercussions for the linguistic practices of the community at large. At first sight, potential repercussions may appear to be negligible. Typically, only one or two members of the community work on transcription in close collaboration with the documentation team. The rest of the community would appear not to be affected by it and is usually also not very interested in what is for them an obscure activity. However, we do not know for sure that this is indeed the case. To date, what the community at large actually knows about the transcription process and whether transcription can influence the overall **language ecology**, however subtly, has never been investigated.


In this regard, it would do well to remember that the production of written language materials—often a dictionary or a reading primer—obviously changes the language ecology in all those instances where the linguistic variant in question was not represented in writing before the documentation project started. The production of such materials involves standardization on many levels: determining base forms for lexical entries, orthographic conventions (e.g., how clitics are written), and so on. Many of these decisions are also part of the transcription process. Usually no attempt is made to include a larger group of community members in these decisions at the transcription stage. But it stands to reason that in one way or another they will indeed affect the community, given that it is a basic goal of modern language documentation to be available and accessible to the community. Hermes & Engman (2017) provide a very instructive example of how the inclusion of a larger group of speakers in transcription changes the documentation process, making it more accessible and relevant for the speech community (in this case, for revitalization).

5. Conclusion It is only a minor exaggeration to say that language documentation is all about transcription. The major argument for a separation of documentation from description—on the conceptual level, not in actual practice where such a separation is impossible—has always been that such a separation helps to focus on those aspects of linguistic fieldwork and language description that tended to be overlooked in descriptive linguistics as practiced throughout the 20th century. Transcription is at the very center of these overlooked and underscrutinized aspects, and documentary linguistics has not yet properly engaged with this core challenge in its subject matter. The brief remarks offered here hopefully make clear not only the importance of properly investigating the transcription process, but also the potential and promises of such investigations for understanding language and the human mind.

References

- Akinnaso, F. Niyi. 1982. On the differences between spoken and written language. *Language and Speech* 25. 97–125.
- Akinnaso, F. Niyi. 1985. On the similarities between spoken and written language. *Language and Speech* 28. 323–359
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, Douglas. 2014. Using multidimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14. 7–34.
- Bucholtz, Mary. 2007. Variation in transcription. *Discourse Studies* 9. 784–808.
- Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*, 35–53. Norwood: Ablex.
- Chafe, Wallace. 1994. *Discourse, consciousness, and time*. Chicago: University of Chicago Press.
- Coupland, Nikolas. 2007. *Style: Language variation and identity*. Cambridge: Cambridge University Press.
- Crowley, Terry. 2007. *Field linguistics: A beginner's guide*. Oxford: Oxford University Press.
- DuBois, John W., Stephan Schuetze-Coburn, Danae Paolino & Susanna Cumming. 1992. *Discourse transcription*. Santa Barbara Papers in Linguistics 4. Santa Barbara, CA: Department of Linguistics, University of California, Santa Barbara.
- Edwards, Jane & Martin D. Lampert (eds). 1993. *Talking data: Transcription and coding in discourse research*. Hillsdale: Lawrence Erlbaum.
- Hermes, Mary & Mel M. Engman. 2017. Resounding the clarion call: Indigenous language learners and documentation. In Wesley Y. Leonard & Haley De Korne (eds.), *Language Documentation and Description*, vol 14. London: EL Publishing, 59–87. <http://www.elpublishing.org/PID/152>
- Himmelmann, Nikolaus P. 2006. The challenges of segmenting spoken language. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 253–274. Berlin: Mouton de Gruyter.
- Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk & Volker Unterladstetter. 2018. On the universality of intonational phrases in spontaneous speech – a cross-linguistic interrater study. *Phonology* 35. 207–245.
- Jung, Dagmar & Nikolaus P. Himmelmann. 2011. Retelling data: Working on transcription. In Geoffrey Haig, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.), *Documenting endangered languages*, 201–220. Berlin: Mouton de Gruyter.
- Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: The University of Pennsylvania Press.
- Ladd, D. Robert. 2014. *Simultaneous structure in phonology*. Oxford: Oxford University Press.
- Maas, Utz. 2009. Orality vs. literacy as a dimension of complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 164–177. Oxford: Oxford University Press.
- Maas, Utz. 2010. Literat und orat. Grundbegriffe der Analyse geschriebener und gesprochener Sprache. *Grazer Linguistische Studien* 73. 21–150. <http://unipub.uni-graz.at/gls/periodical/pageview/1276544>

- Marten, Lutz & Malin Petzell. 2016. Linguistic variation and the dynamics of language documentation: Editing in 'pure' Kagulu. In Mandana Seyfeddinipur (ed.), *African language documentation: New data, methods and approaches* (Language Documentation & Conservation Special Publication 10), 105–129. <http://hdl.handle.net/10125/24651>
- Mosel, Ulrike. 2014. Putting oral narratives into writing – experiences from a language documentation project in Bouganville, Papua New Guinea. In Bernard Comrie & Lucía Golluscio (eds.), *Language contact and documentation / Contacto lingüístico y documentación*, 321–342. Berlin: De Gruyter.
- Nagy, Naomi & Devyani Sharma. 2013. Transcription. In Robert J. Podesva & D. Devyani Sharma (eds.), *Research Methods in Linguistics*, 235–256. Cambridge: Cambridge University Press.
- Ochs, Elinor. 1979. Transcription as theory. In Elinor Ochs & Bambi B. Schieffelin (eds.), *Developmental pragmatics*, 43–72. New York: Academic Press.
- Seifart, Frank, Nicholas Evans, Harald Hammarström & Stephen C. Levinson. 2018. Language documentation 25 years on. *Language* 94, e324-e345. DOI: 10.1353/lan.2018.0070
- Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg R. Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzluft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock & Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 10. 353–402. <http://www.gespraechsforschung-ozs.de>.
- Verschik, Anna. 2015. Language contact, language awareness, and multilingualism. In Jasone Cenoz, Durk Gorter & Stephen May (eds.), *Encyclopedia of language and education*. Dordrecht: Springer. doi: 10.1007/978-3-319-02325-0_21-1.

Nikolaus P. Himmelmann
sprachwissenschaft@uni-koeln.de
 orcid.org/0000-0002-4385-8395

Why cultural meanings matter in endangered language research

Lise M. Dobrin
University of Virginia

Mark A. Sicoli
University of Virginia

In this paper we illustrate why it is important for linguists engaged in endangered language documentation to develop an analytical understanding of the cultural meanings that language, language loss, and language documentation have for the communities they work with. Acknowledging the centrality of cultural meanings has implications for the kinds of questions linguists ask about the languages they are studying. For example: How is age interpreted? What reactions are provoked by accented speech or multilingualism? Is language shift experienced as a painful loss, or a source of newfound freedom, or both? It affects the standards we set for what counts as a satisfying explanation for language endangerment, with prediction necessarily limited in sociogeographic scope. It has implications for the research methods employed, calling for serious engagement with the particular histories and interpretive practices of local linguistic communities. Analyzing cultural meanings can help us see how language use and changes in language use are experienced and therefore acted on by people whose communicative behavior we are concerned with. It can help us interpret why language shift is taking place in a particular community, guide the practices of language documentation and preservation that linguists engage in with that community, and contribute to effective revitalization.

1. Introduction In this paper we illustrate why it is important for linguists engaged in endangered language documentation to develop an analytical understanding of the cultural meanings that language, language loss, and language documentation have for the communities they work with. A number of publications make the claim that ethnography has a prominent role to play in linguistic field research, both as a form of knowledge and as a method that can lead to that knowledge (Hymes 1971[1962]; Harrison 2005; Hill

2006; Ahlers 2009; Dobrin 2008; Dobrin and Berson 2011; Sicoli 2011; Childs, Good, and Mitchell 2014; Di Carlo 2016; Dobrin and Schwartz 2016). Yet the basic message of these publications—that cultural factors cannot be treated as an externality to documentary linguistics without compromising both the research process and its outcomes—has not had a great impact on the field. For example, the message does not generally hold a prominent place in linguistic field methods courses or other kinds of training given to linguistics students preparing for fieldwork. Whereas discourses of research ethics and community collaboration have become ubiquitous and students of documentary linguistics are now systematically conversant in them, students are not being similarly introduced to participant observation as a method that can contribute to the development of cultural understanding, or being urged to familiarize themselves with the major themes in the anthropological literature on the areas where they plan to work.¹In this respect, there is still room for growth in the field of documentary linguistics that Himmelmann's pivotal (1998) publication helped create. Himmelmann (2008: 338) recognized this as a problem for the field:

For most research areas of concern to core linguistics, e.g., grammatical theory or typology, it is not clear to what extent the disregard for social aspects of language structure and use compromises research goals and outcomes. However, this disregard is indeed harmful to a number of topic areas. One of these areas is large-scale language endangerment....

[T]he essentially a-social conceptualization of linguistic knowledge within mainstream structural linguistics... has delegated to the subfield of sociolinguistics (broadly conceived, including anthropological linguistics) the investigation of all social aspects of language structure and use. In putting language endangerment on the mainstream agenda, structural linguistics has added another issue to the growing list of items that second guess the wisdom of excluding from its core agenda almost all regard for the ways in which linguistic knowledge is socially constructed and reproduced.

So in this paper we lay out a number of ways in which cultural meanings matter for the study of endangered languages, from interpreting why language shift is taking place in a particular local community, to helping guide the practices of language documentation and preservation that linguists engage in with that community, to planning for effective revitalization.

Acknowledging the centrality of cultural meaning has implications for the kinds of questions linguists ask about the languages they are studying. For example: How is age interpreted? What reactions are provoked by accented speech or multilingualism? Is language shift experienced as a painful loss, or a source of newfound freedom, or both? It affects the standards we set for what counts as a satisfying explanation for language endangerment, with prediction necessarily limited in sociogeographic scope. It also has implications for the methods employed, calling for serious engagement with the particular histories and interpretive practices of local linguistic communities (Di Carlo 2016; Di

¹Collaborating with researchers who have different disciplinary skill sets may also be desirable, but it cannot take the place of developing one's own understanding of key cultural themes that connect language with other domains of social life in an intended area of fieldwork, as these will affect language use and distribution, undergird patterns of shift, and have implications for revitalization interventions. Entering a community as a fieldworker comes with personal responsibility.

Carlo and Good 2017; Good and Di Carlo in press; Lüpke and Storch 2013). This includes, but goes beyond, uncovering the attitudes and beliefs people have about language—the linguistic ideologies they hold—in the sense of associations with whole codes such that people find them good to speak or not speak. It also means understanding the cultural mechanisms that help construct those ideologies and hold them in place within webs of meaning and action (see, e.g., Schieffelin, Woolard, and Kroskrity 1998; Irvine and Gal 2000). Ideologies are more than series of associations; they are like the grammar of social life that tie patterns of activity, ideas, and affect together. And they are often unconscious: there may be indicators in the things people say, but they will rarely arise as the answers to overt questions. Analyzing these kinds of meanings can help us see how language use, and changes in language use, are experienced and therefore acted on by the people whose communicative behavior we are concerned with.

2. Shift There have been a number of efforts to describe and analyze the factors and forces involved in language shift, culminating recently in a call for linguists to develop a generalized predictive model of “why and how some languages become endangered, die, survive threats to them, or even thrive” (Mufwene 2017: e202, n.2). Yet the substantial literature already directed toward this end seems to have reached a limit: Pauwel’s (2016) textbook *Language Maintenance and Shift* winds up its assessment of the current state of knowledge about language shift dynamics by saying,

the majority of factors that have been discovered and examined in relation to [language maintenance] or [language shift] seldom have the same impact across... settings or linguistic groups.... As a result, [we have] not yet been able to come up with a convincing model or theory that can predict, reliably, which factors or combinations of factors lead to a specific outcome. (98)

This is borne out by two of the articles responding to Mufwene’s call, which diverge in their assessment of something as basic as speaker numbers, with Bower (2017) citing work that upholds the generalization that small population size is a driver of language shift, and Lüpke (2017) arguing forcefully that in an African context it is not.

In approaching the question of how and why shift takes place linguists have tended to rely on categories like speaker numbers, domains of use, utility for employment, etc. that can be treated as independent variables. Yet there is an almost unimaginable range of ways in which language can be refracted through cultural categories and practices. This means that even common factors will not always be commensurable. Take, for example, something as seemingly straightforward as age distribution. When working with Kaska teenagers in the Yukon, Meek (2007) found that their association of the language with authority had become so strong that it shaped both what teens would say in the language (they used it especially for directives) and their understanding that elders were the only social group that properly spoke the language, which reinforced shift. When working with bilingual indigenous Mixe speakers, Suslak (2009) found that code switching with Spanish held different meanings across generations, with youth spurning their parents’ mixing of the two languages as sloppy and careless, while they expressed their own sophistication by attending to the language boundary so carefully that it led them to hypercorrect. Treating age as a variable that can be straightforwardly compared across settings misses the way language shift is shaped by these kinds of local meanings.

Scholars going back at least to the 1960s have been producing “lists, typologies or taxonomies of factors and variables” (Pauwels 2016: 105) meant to explain the dynamics

of language maintenance vs. shift (Ferguson 1962; Stewart 1968; Haugen 1972; see also Campbell 2017). Probably the most comprehensive of these is presented in Grenoble and Whaley (1998), which builds on prior work by John Edwards (1992). Grenoble and Whaley identify economics as “the single strongest force influencing the fate of endangered languages” and attribute to it “the potential...to outweigh all others combined” (1998: 52, 31). Yet they also acknowledge that “it is at the level of micro-variables where one can account for how differences in the rate, outcome, and reversibility of language-shift cases come about” (1998: 28). Micro-variables, “characteristics which are unique to specific speech communities,” cannot be entirely equated with cultural factors, but they often have a cultural component. For example, the impact of literacy in a community must take account of its “*social meaning...*, a set of micro-variables which involve the attitudes, beliefs, and values of a community” (Grenoble and Whaley 1998: 33). Just what kinds of “social meanings” might be involved? Here lies the problem with any “ready typology of language shift that we can apply consistently across cultures” (Sicoli 2011: 163). Preselecting the categories deemed to be relevant—no matter how expansive—limits our ability to learn how language shift is structured and experienced in a given local situation (Dobrin 2010).

Consider the transmission of Eastern Tukanoan languages in the multilingual Vaupés region of the northwest Amazon described by Janet Chernela and others. In this part of the world, language loss and maintenance are both going on at once in the same households as the ordinary state of affairs. In fact, Chernela (2004: 13) explicitly compares the language situation in the Vaupés to that of immigrants in the U.S. whose children know their home languages but grow up to not use them. The Vaupés culture area is known for its linguistic exogamy: people marry outside their own language group, with marriage to a fellow speaker held to be incestuous because language is culturally construed as an embodied substance passed down through descent (Chernela 2018). This practice, along with patrilocal residence, results in a situation where children are raised in bilingual households but nevertheless become monolingual speakers of their father’s language. At the same time, knowledge of the mother’s language is suppressed and stigmatized, although not forgotten, making this a place in which people are monolingual speakers, but bilingual hearers. Moreover, “every attempt is made to avoid hybridization, since it is considered essential that linguistic identities remain distinct and linguistic boundaries be kept stable” (Chernela 2004: 15). In the social-symbolic configurations created through these cultural practices, monolingualism is associated with men, whereas women often end up speaking their parents’ language peripherally, when at home with their children and with fellow in-married women who happen to hail from their original language group. Monolingualism is culturally elaborated as a display of self-discipline associated especially with males, whereas multilingualism and code-switching are felt to be feminine, chaotic, and politically destabilizing. Lapses in monolingual self-control are thus cause for humiliation and shame in men, leading them to deny that they ever code-switch, even though they sometimes do so in order to facilitate communication with their in-laws. As children begin producing their first utterances, their mothers guide them away from their own language and instead toward the father’s by use of feigned incomprehension or outright correction, making the process of language learning, in Chernela’s words, “an early form of mother-separation” (2004: 19). In its preferred form, marriage is to the child of a mother’s brother, so later in life there is often a return to the mother’s language in speech between spouses, which transforms the suppressed language of mother-infant intimacy into “the language of affect and libido” (Chernela 2004: 19). In short, while the

linguistic outcome for individuals might be similar to what we find among children of immigrants shifting to English in the U.S., the cultural bases for the outcome are all but incomparable and could hardly have been imagined without in-depth ethnographic study.

Cultural meanings can remain constant even as language change takes place, so that they form part of the logic that organizes shift as a social process. Perhaps the best-known example of this is Kulick's (1992) study of language shift in Gapun village in the New Guinea Sepik. Another example comes from Dobrin's work on coastal varieties of Mountain Arapesh in Papua New Guinea, which are now spoken almost exclusively by elders and may soon cease to be spoken at all. In Arapesh communities, as elsewhere in Melanesia, many high-status cultural forms are acquired from elsewhere, rather than originated group-internally, with the social distance traversed in order to acquire them contributing to their value. In part this follows from difficulties of mobility due not to the renowned ruggedness of the New Guinea landscape but to the vulnerability associated with traveling across other peoples' lands. Safe movement beyond one's home locality was traditionally structured according to *roads*, which represent both real, physical pathways and series of inter-locality relationships, so that the further out along the road one went from home the more social capital in the form of "road friends" or allies one could be inferred to have. It was therefore an Arapesh cultural disposition to associate the self with markers of foreignness; leading one early ethnographer (Mead 1938) to describe Arapesh as "an importing culture". In Melanesia, languages as codes and high-status speech genres such as songs, magic spells, and oratory are among the cultural forms that acquire value through this partly practical, partly symbolic system, which has been convincingly argued to have reflexes in grammatical structure (Aikhenvald 2007). But with the cessation of warfare brought about by colonial control and the subsequent formation of an encompassing state, the obstacles to travel and thus cultural importation are no longer in place, so that the English-based contact language Tok Pisin, which is associated with Europeans and hence the greatest possible social distance, can be—and has been—readily learned by all (Dobrin 2014: 143). When the state-based hierarchization of language (described below) was superimposed upon this acquisitive ethos, it led to the rapid and dramatic shift to Tok Pisin that has taken place throughout the region.

The cultural symbolic drivers of shift can also change over time: what causes language loss may be quite independent from what motivated and sustained multilingualism beforehand. This means that shift can not only mean different things across cultures, but within them over the time course, even in adjacent generations. An example is the case of Mexicano (Nahuatl) in Central Mexico. After centuries of coexistence with Spanish, a reinterpretation of Mexicano-Spanish bilingualism in the later 20th century led the next generation to shift. In their 1986 book, *Speaking Mexicano*, Jane and Kenneth Hill show how Spanish code-switching and borrowing changed in meaning when a new generation of middle-aged Mexicano men latched onto it as a way to challenge the authority of their elders, who displayed their power and knowledge through their control of Spanish forms. This intergenerational dynamic was reflected in the discourse around code switching and borrowing, leading to a reinterpretation of the use of Spanish forms as "contamination". The purist movement that resulted raised the bar for speaking the native language so high that the youth dared speak only Spanish, lest they be subject to sanction for their "impure" Mexicano.

With the centrality of cultural meanings in mind, we can revisit what is often considered the single greatest influence on language vitality, economics, with an acknowledgment that it, too, is symbolically mediated:

[C]ompared to the majority/dominant population, local community members are relatively powerless politically, and are less educated, less wealthy... with less access to modern conveniences and technologies.... [T]his socially disadvantaged position becomes associated with... the local language and culture, and so knowledge of the local language *is seen as* an impediment to social and economic development. Socioeconomic improvement thus comes to be *perceived as* tied to knowledge of the language of wider communication, coupled with renunciation of the local language and culture. (Grenoble 2011: 34, emphases ours)

People's linguistic practices become bound up with the unifying hierarchy of the state, such that linguistic differences "cease to be incommensurable particularisms" and instead come to be interpreted as inferior deviations from legitimate or standard forms of speech (Bourdieu 1982: 54). Imagine a cone on a three-dimensional graph: the further some form of linguistic expression diverges from the standard-language center, the further it falls on the scale of value (Silverstein 2017: 135). It is this whole cultural system, which Dorian (1998), following Grillo (1989), calls an "ideology of contempt" for non-dominant languages, that has been exported by Europeans throughout the world along with their standardized languages at the top. The symbolic nature of even economically motivated shift is demonstrated by how often "marginalized groups remain marginalized" even after they shift: "There is no convincing evidence that the shift to another language or repertoire yields real—as opposed to imagined or desired—socioeconomic advantages. These ideas operate at the ideological level... [and] are in many contexts not grounded in real economic gains" (Lüpke 2015: 72).

3. Documentation and Preservation Because the research process often involves direct, intense, instrumental interaction that brings people together across cultures, the way cultural meanings bear on language documentation and preservation are too numerous to adequately survey. They range from how the technologies of writing and recording are understood, to notions about the relation between self and outsider that are brought to the fore in both linguistic fieldwork and language shift, to the assessment of what constitutes an appropriate or authoritative speaker. As with the influences on language shift, the meanings that will be relevant in a particular situation cannot be presupposed, but they can be discovered through attentive observation as the research process unfolds. Here we present just a handful of cases that illustrate how local understandings can influence language documentation as an activity, as well as its products.

While documenting the Teop language of Bougainville, Papua New Guinea, Mosel (2015) found that local research assistants editing transcripts of recorded legends for community distribution were creating an entirely new linguistic register that drew not just on project goals but on local ideas of what writing or storytelling should be like. For example, they often made constructions more complex by explicitly marking the links between clauses or by combining them into a single clause. Similarly, while archiving a set of Bukiyip Arapesh texts that had been collected by another linguist and transcribed by native speakers in the 1970s, Dobrin (2017) discovered that the transcripts diverged from the audio recordings they were based on in numerous ways. Borrowed elements were replaced with their vernacular equivalents, obscuring the extent to which the influence of Tok Pisin had already advanced at that time; canonical Melanesian discourse forms

such as tail-head linked structures had their redundancy removed, and turns by non-focal participants were left untranscribed despite being critical for understanding the meaning of the text. Departures such as these result in something other than “documentation” as that term is now understood, but as Mosel (2015) points out, they do offer an interesting new angle from which to explore local understandings of the relation between spoken language and writing.

The responses of linguistic consultants may reflect their habituation to methods used in prior research projects, or to parallel-feeling activities like classroom instruction or local socialization practices. For example, where prior research has involved eliciting translations of sentences through a contact language, speakers may understand the goal of linguistic research to *always* be the provision of such translations. When Sicoli began building a video corpus of spontaneous Zapotec interactions after several years of collecting texts, eliciting vocabulary, and conducting psycholinguistic experiments, he found that the linguistic consultants he had worked with on the prior projects had to go through a period of adjustment to the new work style (Sicoli ms.). Community responses to the project were also shaped by people’s past experiences with language work. While many appreciated having a video corpus that showed the language in use in everyday life, others, who had learned to value the more formal kind of speech that is produced through elicitation, commented that there were “better” examples of Zapotec to be found than the ones in the recordings. Sicoli and Kaufman’s (2017) use of a standard survey instrument to elicit Zapotec and Chatino-language utterances through Spanish prompts offers another example of how the documentation process can be shaped by local participants’ understandings. When elder speakers worked with younger interviewers, they sometimes responded with imperative forms regardless of what inflection was implied by the prompt, as the situation seemed to them like an appropriate one in which to express their linguistic authority by using the language to tell the interviewer what to do. These speakers were responding not just to the prompts, but to the wider social configuration in which the research was taking place, as they interpreted it.²

Who counts as a native speaker for purposes of language work, and when and where they consider it appropriate to inhabit the speaker’s role, is another area where local understandings can challenge linguists’ assumptions. It might be preferable from the linguist’s point of view to work with speakers who command a wide range of registers, have the ability to express themselves using complex constructions, and exhibit minimal interference from the phonology and vocabulary of a contact language. But the community may prioritize other considerations. Myaamia language activist Daryl Baldwin said that being “able to explain what they were saying with some cultural context” was a more important criterion for speakerhood than being able to “hold extended conversation... in the language”; similarly, Warm Spring Language Program Director Myra Johnson said, “If they speak a broken Native American language, then maybe that’s how they learned it”, so it should not preclude them from serving as a linguistic expert (Leonard and Haynes 2010: 285). Evans (2001) writes about the challenge of determining who is a speaker of Australian Aboriginal languages, where linguistic authority or “ownership” is based not on fluency but on affiliation with the kin group on whose lands the language was traditionally spoken. He also shows how speakerhood

²Briggs 1986 offers an extended argument that the format of “the interview” (including of course the linguistic elicitation interview) is too often taken for granted, when it is actually shaped by participants’ interpretations of what is happening as a communicative event. These interpretations have implications for the social roles the participants inhabit, the speaking styles they use, the interactional goals they aim to achieve, etc.

can change qualitatively when, for example, someone with a greater right to language ownership passes away, leaving another speaker closer than they had been to the center of linguistic authority, or when what he calls an “amplifier” is present. An amplifier may have only partial skills, but the presence of such a person may draw out others who have greater fluency but would be unable or hesitant to use the language on their own.

When conducting documentary linguistic fieldwork on the Yopno language in Papua New Guinea, Slotta (2015) found that he was repeatedly being offered the same narrative to record: a story about an American who had cared for a member of his host family during WWII. Recording the same story over and over seemed like a waste of scarce battery power and did not serve the goals of a project that was meant to document speech across a range of topics and genres. But it eventually dawned on Slotta that the story was not being offered to serve the documentary record. Rather, it was being told to justify his close association with one particular household in the village, the one in which he lived, since this arrangement skewed his exchange relations to that family’s benefit. From the speakers’ point of view, repeating this story was not inefficient language documentation but a strategy for maintaining community harmony by justifying the present situation through reference to past events. Local understandings of the researcher’s presence as a person cannot be separated from the research process, and can even give shape to the documentary material collected. Slotta’s Yopno interlocutors also enthusiastically told him stories that are not normally shared outside of clans and lineages, and many of the storytellers were adamant that these secret clan histories should be made public in digital archives. They did this because they believed computers had the power to reveal hidden knowledge and so might confirm or even fill in missing details, which could support their claims to land. So participants in the documentation project were pursuing their own interests according to their own understanding of how the world works, which had little to do with the linguist’s understanding or the goal of preserving language and culture. This case shows how documentation of situated speech may simultaneously document an encounter between cultural perspectives.

In his 1998 article about “the Yellowman Tapes,” folklorist Barre Toelken explains his decision not to preserve the collection of audio recorded Navajo texts he had made over the course of his 30-year career. The concern was not his rights over the recordings, but the potential hazards of the powerful speech they capture when one cannot be sure where or when it will be respoken. In Navajo linguistic cosmology speech does not just describe reality, it creates it: the release of utterances into the air is understood to act directly on nature, including on the spirits of living beings. So the Coyote stories featured most prominently in Toelken’s recordings—which are felt to be edifying precisely because they dramatize problematic or inappropriate scenarios—could result in a dangerous disequilibrium if they were replayed in the wrong circumstances, for example out of season.³ Moreover, replaying the recordings would now bring hearers in contact with the voices of the dead, something many Navajos try to avoid. So while Toelken recognized the unfortunate loss to knowledge that his decision entailed, he felt the only justifiable course of action was to accede to his Navajo interlocutors’ wishes and return the recordings to them rather than preserve them in an archive.

Our final illustration of the cultural complexities of documentation and preservation comes from Blumenthal’s (2011) work in the Loba community of Lo Monthang, Nepal, where she recorded a repertoire of historically significant ritual songs called Garlu. The

³Toelken 1996 discusses how his own repeated playing of the recordings over his research and teaching career had precisely this effect, bringing physical harm and even loss of life to some of his close Navajo associates.

musicians who sang these songs were originally Muslims who had long been incorporated into this Buddhist community, but they continued to handle instruments like cowhide drums that were untouchable by others, and these practices were seen to justify their place in the lowest caste of the social hierarchy. In part because of their lowly, even shameful social status, the musician class was dwindling as sons opted to leave their remote community to work as contract laborers for the U.S. military in Iraq and Afghanistan or join the Maoist party rather than take up their inherited place as musicians at the center of their village's ritual functioning. The Loba diaspora living in New York were nostalgic about their heritage and concerned about its loss, so they were grateful that Blumenthal had agreed to work with a local musician in Lo Monthang to make CDs and a songbook they could distribute and listen to remotely. The primary village musician appreciated the positive attention and welcomed the opportunity to document his songs. But Blumenthal found that her work was creating problems back in the village. If the actual musicians were expendable, it only further justified the wider community's derisive attitude toward them. At the same time, it created anxiety among the villagers about the continuity of their way of life, because the one expert musician who remained was still regularly playing in rituals that were considered necessary for the whole community's purity. So in this situation, the activity of documenting and preserving the Garlu repertoire was difficult to dissociate from the social position of the musicians who performed it, local spiritual beliefs and practices, and the changing social structure and geographic dispersal of the community.

4. Revitalization If symbolically organized cultural meanings play a role in language shift and documentation, then it should not come as a surprise that they also have an important role to play in language revitalization, something also noted in Leonard's (2017: 19) call for linguists to open themselves to "Indigenous definitions of 'language'". The following examples make clear how this is so.

For the contemporary Garifuna community in Guatemala studied by Alison Broach, shift away from Garifuna is experienced as disruptive to communal harmony because it cuts people off from their dead ancestors, who continue to participate in social life by advising and reprimanding their descendants in Garifuna through dreams and ritual trances. This moral imperative for young people to listen to ancestors' voices has in turn influenced community efforts to address the problem of language shift as they experience it. Revitalization workshops are configured like spirit possession rituals, with elders conversing with youth and offering them guidance in Garifuna in a familial setting, just as the dead do when they ritually connect with their living kin. As Broach (2017) points out, having a culturally significant population of speakers *who are also dead* adds a whole new dimension of complexity to the problem of assessing speaker numbers.

Josh Wayt's work with members of the Dakota community at Lake Traverse Reservation similarly points to the primacy of moral meanings in regard to changing language practices. Why exactly are Lake Traverse residents so invested in revitalizing their native language? After all, this is a community coping with seemingly more pressing social problems like drug abuse and violence, and its culturally distinctive identity is strong even without the language because of the continued vibrancy of spiritual and ceremonial life. What Wayt (2018) observes is that in Lake Traverse, elders' (and, more generally, teachers') talk about the Dakota language—down to lessons on grammatical patterns like morpheme order—is loaded with references to moral relations with kin. This is in line with traditional methods of teaching and correcting, which tend to be highly

indirect. In other words, the local investment in the native language *is* actually targeting the most pressing social problems by scaffolding a discourse about moral relations within the community. To be considered successful in this setting, language revitalization will entail so much more than the acquisition of language skills. In fact, what it entails is so different that Leonard (2017) proposes a separate name for it: not “revitalization” but “reclamation”.

5. Conclusion We have aimed to show here that documentary linguistics and related practical efforts like language preservation and reversing language shift that take “action on language” (Costa 2016: 2) must recognize the meanings language has for local actors, as these inevitably form the backdrop, if not the foreground, for such efforts. What makes this a challenge is that cultural meanings—as with those that guide linguists’ own goals and behaviors—may not be readily articulable, and so cannot necessarily be queried through interviews or similarly direct methods. Moreover, local cultural meanings interconnect language with other domains like kinship, gender, age, spiritual beliefs, morality, and so on, creating webs of symbolic relations that can be hard to disentangle from one another and from those features that seem more obviously to be “about language”. But linguists’ assumptions about where the limits of language are, or ideas about what the goals and effects of language documentation work should be, may or may not align with those held by those most immediately involved. Thus, reading available ethnographic work on the relevant region and integrating research practices like participant observation into language documentation can both help researchers respond appropriately to the wider ethnolinguistic scene and refrain from reproducing boundaries between domains that are only artifacts of the historical division separating linguistics from other disciplines since Saussure. In the twenty years since Himmelmann first proposed that linguists construe language documentation as its own subfield of linguistics there has been talk about forging such connections between disciplines, but perhaps because institutions are so reified and disciplinary cultures slow to change, linguistics has still not fully taken to heart the lessons of anthropological ethnography that foster exploration of the patterns that connect seemingly disparate domains. Finding ways to overcome the artificial and unhelpful boundaries between linguistics and anthropology remains a continuing challenge.

References

- Ahlers, Jocelyn C. 2009. The many meanings of collaboration: Fieldwork with the Elem Pomo. *Language & Communication* 29(3). 230–243.
- Aikhenvald, Alexandra Y. 2007. Multilingual fieldwork, and emergent grammars. BLS 33(1). 3–17. <http://dx.doi.org/10.3765/bls.v33i1.3513>
- Blumenthal, Katharine A. 2011. *Hierarchical transformation and song circulation: Documenting the Garlu folksongs of Lo Monthang, Nepal*. Charlottesville, VA: University of Virginia. (Unpublished MA thesis.)
- Bourdieu, Pierre. 1982. *Language and symbolic power*. Cambridge, MA: Harvard University Press.
- Bowern, Claire. 2017. Language vitality: Theorizing language loss, shift, and reclamation. *Language* 93(4). e243–e253.
- Briggs, Charles L. 1986. *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge: Cambridge University Press.
- Broach, Alison. 2017. The voice of the nation: Garifuna ancestors as political actors in language. Washington, D.C. (Poster presented at the American Anthropological Association Annual Meeting.)
- Campbell, Lyle. 2017. On how and why languages become endangered. *Language* 93(4). e224–e233.
- Chernela, Janet M. 2004. The politics of language acquisition: Language learning as social modeling in the Northwest Amazon. *Women & Language* 27. 13–21.
- Chernela, Janet M. 2018. Language in an ontological register: Embodied speech in the Northwest Amazon of Columbia and Brazil. *Language & Communication* 63. 23–32.
- Childs, Tucker, Jeff Good, and Alice Mitchell. 2014. Beyond the ancestral code: Towards a model for sociolinguistic language documentation. *Language Documentation & Conservation* 8. 168–191.
- Costa, James. 2106. *Revitalising language in Provence: A critical approach*. Malden, MA: Wiley-Blackwell.
- Di Carlo, Pierpaolo. 2016. Multilingualism, affiliation and spiritual insecurity: From phenomena to processes in language documentation. *Language Documentation & Conservation* 10. 71-104 <http://hdl.handle.net/10125/24649>
- Di Carlo, Pierpaolo and Jeff Good. 2017. The vitality and diversity of multilingual repertoires. *Language* 93(4). e254–e262.
- Di Carlo, Pierpaolo, Jeff Good & Rachel Ojong.. In press. Multilingualism in rural Africa. *Oxford Research Encyclopedia of Linguistics*.
- Dobrin, Lise M. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment from Melanesia. *Language* 84(2). 300–324.
- Dobrin, Lise M. 2010. Review of *Language and poverty*, ed. by Wayne Harbert, Sally McConnell-Ginet, Amanda Miller, and John Whitman. *Language Documentation & Conservation* 4. 159–168.
- Dobrin, Lise M. 2014. Language shift in an “importing culture”: The cultural logic of the Arapesh roads. In Peter K. Austin & Julia Sallabank (eds.), *Endangered languages: Beliefs and ideologies in language documentation and revitalization*, 125–148. Proceedings of the British Academy 199. London: Oxford University Press.
- Dobrin, Lise M. 2017. The ever-deepening meaning of the Bukiyip Arapesh suitcase miracle. Washington, D.C. (Presentation in session on The Social Lives of Linguistic Legacy Materials, American Anthropological Association Annual Meeting)


- Dobrin, Lise M. & Josh Berson. 2011. Speakers and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 187–211. Cambridge: Cambridge University Press.
- Dobrin, Lise M. and Saul Schwartz. 2016. Collaboration or participant observation? Rethinking models of “linguistic social work”. *Language Documentation & Conservation* 10. 253–277.
- Dorian, Nancy. 1998. Western language ideologies and small-language prospects. In Lenore A. Grenoble & Lindsay J. Whaley (eds.), *Endangered languages: Language loss and community response*, 3–21. Cambridge: Cambridge University Press.
- Edwards, John. 1992. Sociopolitical aspects of language maintenance and loss: Towards a typology of minority language situations. In Willem Fase, Koen Jaspaert & Sjaak Kroon (eds.), *Maintenance and loss of minority languages*, 37–54. Amsterdam: John Benjamins.
- Evans, Nicholas. 2001. The last speaker is dead—Long live the last speaker! In Paul Newman & Martha Ratliffe (eds.), *Linguistic fieldwork*, 250–281. Cambridge: Cambridge University Press.
- Ferguson, Charles A. 1962. The language factor in national development. *Anthropological Linguistics* 4(1). 23–27.
- Grenoble, Lenore A. 2011. Language ecology and endangerment. In Peter K. Austin & Julia Sallabank (eds.), *Handbook of endangered languages*, 27–44. Cambridge: Cambridge University Press.
- Grenoble, Lenore A. and Lindsay J. Whaley. 1998. Toward a typology of language endangerment. In Lenore A. Grenoble & Lindsay J. Whaley (eds.), *Endangered languages: Language loss and community response*, 22–54. Cambridge: Cambridge University Press.
- Grillo, Ralph D. 1989. *Dominant languages: Language and hierarchy in Britain and France*. Cambridge: Cambridge University Press.
- Harrison, K. David. 2005. Ethnographically informed language documentation. In Peter K. Austin (ed.), *Language Documentation and Description* 3. 22–41. London: SOAS.
- Haugen, Einar. 1972. *The Ecology of Language*, ed. by Anwar S. Dil. Stanford, CA: Stanford University Press.
- Hill, Jane H. 2006. The ethnography of language and language documentation. In Jost Gippert, Nikolaus P. Himmelmann, & Ulrike Mosel (eds.), *Essentials of language documentation*, 113–128. Berlin: Mouton de Gruyter.
- Hill, Jane H. and Kenneth C. Hill. 1986. *Speaking Mexicano: Dynamics of syncretic language in central Mexico*. Tucson: University of Arizona Press.
- Himmelmann Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–95.
- Himmelmann, Nikolaus P. 2008. Reproduction and preservation of linguistic knowledge: Linguists’ response to language endangerment. *Annual Reviews in Anthropology* 37. 337–350.
- Hymes, Dell. 1971[1962]. The ethnography of speaking. In Thomas Gladwin & William C. Sturtevant (eds.), *Anthropology & Human Behavior*, 13–52. Washington, D.C.: The Anthropological Society of Washington.
- Judith T. Irvine & Susan Gal. 2000. Language ideology and linguistic differentiation. In Paul V. Kroskrity (ed.), *Regimes of Language: Ideologies, politics, and identities*, 35–83. Santa Fe: School of American Research Press.

- Kulick, Don. 1992. *Language shift and cultural reproduction: Socialization, self, and syncretism in a Papua New Guinean village*. Cambridge: Cambridge University Press.
- Leonard, Wesley Y. 2017. Producing language reclamation by decolonising “language”. In Wesley Y. Leonard & Haley De Korne (eds.), *Language Documentation and Description* 14. 15–36. London: EL Publishing.
- Leonard, Wesley Y. & Erin Haynes. 2010. Making “collaboration” collaborative: An examination of perspectives that frame linguistic field research. *Language Documentation & Conservation* 4. 268–293.
- Lüpke, Friederike. 2015. Ideologies and typologies of language endangerment in Africa. In James Essegbey, Brent Henderson, & Fiona McLaughlin (eds.), *Language documentation and endangerment in Africa*, 59–105. Amsterdam: John Benjamins.
- Lüpke, Friederike. 2017. African(ist) perspectives on vitality: Fluidity, small speaker numbers, and adaptive multilingualism make vibrant ecologies. *Language* 93(4). e275–e279.
- Lüpke, Friederike & Anne Storch. 2013. *Repertoires and choices in African languages*. Berlin: De Gruyter Mouton.
- Mead, Margaret. 1938. *The Mountain Arapesh: I. An importing culture*. Volume 36, Part 3. New York: Anthropological papers of the American Museum of Natural History.
- Meek, Barbra A. 2007. Respecting the language of elders: Ideological shift and linguistic discontinuity in a Northern Athapascan community. *Journal of Linguistic Anthropology* 17(1). 23–43.
- Mosel, Ulrike. 2015. Putting oral narratives into writing: Experiences from a language documentation project in Bougainville, Papua New Guinea. In Bernard Comrie & Lucía Golluscio (eds.), *Language contact and documentation – Contacto lingüístico y documentación*, 321–342. Berlin: de Gruyter.
- Mufwene, Salikoko. 2017. Language vitality: The weak theoretical underpinnings of what can be an exciting research area. *Language* 93(4). e202–e223.
- Pauwel, Anne. 2016. *Language maintenance and shift*. New York: Cambridge.
- Schieffelin, Bambi B., Kathryn A. Woolard & Paul V. Kroskrity. 1998. *Language ideologies: Practice and theory*. New York: Oxford University Press.
- Sicoli, Mark A. 2011. Agency and ideology in language shift and language maintenance. In Tania Granadillo & Heidi A. Orcutt-Gachiri (eds.), *Ethnographic contributions to the study of endangered languages*, 161–176. Tuscon, AZ: University of Arizona Press.
- Sicoli, Mark A. Unpublished manuscript. *Saying and doing in Zapotec: Multimodality in the language of joint actions*. University of Virginia.
- Sicoli, Mark A. & Terrence Kaufman 2017. *The survey of Zapotec and Chatino languages collection*. Archive of the Indigenous Languages of Latin America.
- Silverstein, Michael. 2017. Standards, styles, and signs of the social self. *Journal of the Anthropological Society of Oxford* 9(1). 134–64.
- Slotta, James. 2015. Secret stories, public records. Presentation in session on Being there with the language: Language documentation in its ethnographic context. Portland, OR. (Annual Meeting of the Linguistic Society of America.)
- Stewart, William A. 1968. A sociolinguistic typology for describing national multilingualism. In Joshua Fishman (ed.), *Readings in the Sociology of Language*, 531–45. The Hague: Mouton.
- Suslak, Daniel. 2009. The sociolinguistic problem of generations. *Language & Communication* 29(3). 199–209.

- Toelken, Barre. 1996. From entertainment to realization in Navajo fieldwork. In Bruce Jackson & Edward D. Ives (eds.), *The world observed: Reflections on the fieldwork process*, 1–17. Urbana: University of Illinois Press.
- Toelken, Barre. 1998. The Yellowman tapes, 1966-1997. *Journal of American Folklore* 111(442). 381–391.
- Wayt, Josh. 2017. *The voice of our ancestors: Revitalizing Dakota and reconstituting community at Lake Traverse Reservation, SD*. Charlottesville: University of Virginia. (Unpublished dissertation proposal.)


Lise M. Dobrin

ld4n@virginia.edu

 orcid.org/0000-0002-8012-0111

Mark A. Sicoli

marksicoli@virginia.edu

 orcid.org/0000-0002-2658-5700

Reflections on (de)colonialism in language documentation

Wesley Y. Leonard
University of California, Riverside

With origins in colonial logics and institutions, language documentation practices can reinforce colonial power hierarchies and norms in ways that work against the needs and values of Indigenous language communities. This paper highlights major patterns through which this occurs, along with their effects, and models how language documentation can be structured in ways that are more grounded in the experiences and perspectives of the communities that use it. I propose decolonial interventions that emerge from Indigenous research principles and perspectives, and illustrate how these practices can better support language community needs while also improving the scientific value of language documentation.

1. Stories of engagement with language documentation¹ As a linguist who focuses on reversing language shift in Native American communities, much of my professional work involves promoting language documentation, using its resulting products, and engaging with the broader social issues that surround it. For example, I have served many times as an instructor for Breath of Life programs, in which Native Americans access and interpret archival documentation for language reclamation purposes. These programs include training to facilitate the use of legacy documentation, much of which was created by linguists and almost none of which was created for pedagogical purposes.² This process of consulting language documentation to reverse language shift is very close to me since it is what occurred in my own Miami community: Our tribal language, myaamia, was sleeping for about 30 years and was later brought back into community use from archival records (Leonard 2008).

¹I would like to thank Colleen Fitzgerald, Andrea Berez-Kroeker, and an anonymous reviewer for their feedback on an earlier version of this paper. I also offer tremendous gratitude to the participants of the Natives4Linguistics satellite workshop at the 2018 Linguistic Society of America Annual Meeting for sharing their perspectives, which are embedded throughout this paper. Funding for the Natives4Linguistics workshop was provided by the National Science Foundation, BCS grant #1743743.

²See Hinton (2001), Fitzgerald & Linn (2013), and Sammons & Leonard (2015) for details about Breath of Life.

As I also hold a disciplinary interest in language documentation, I share other linguists' concerns about documentation methods, archiving, and the like, and welcome the growing body of literature on these topics with its growing focus on language communities. However, while language community members are increasingly written about, they still rarely serve as primary voices in scholarly outlets. I thus focus this paper on my perspectives as an Indigenous community member, specifically a citizen of the Miami Tribe of Oklahoma, who gained access to myaamia because of documentation. This is a story grounded in lived experiences—my own, and those that have been shared with me by other Indigenous people.

Common to these stories is an emphasis that the (non-)use of a given language emerges from social marginalization that Indigenous people continue to experience. My tribal community shifted away from myaamia almost entirely, and while there are multiple specific causes, the basic underlying theme is *colonialism*. By this, I refer to ideas and practices of subjugation by socio-politically dominant groups or institutions (including academic disciplines) that assert and maintain control over the minds, bodies, and cultures of other groups, generally with an intent of exploiting them to benefit the dominant group. *Decolonialism*, by extension, disrupts the ideas and institutions of colonialism.³ It is a way of thinking and acting that emphasizes the sovereignty, peoplehood, intellectual traditions, and cultural values of groups that experience colonialism. As with the language-specific decolonial movement that I call *language reclamation*, which refers to revitalization efforts that are grounded in and driven by community needs and values (Leonard 2011, 2012, 2017), decolonial approaches in general look not just at a given current situation (e.g., “Language X has only five speakers so we must document it now”) but also at the histories of institutions and power that have fostered it (e.g., “Why does Language X have only five speakers—and whose definition of ‘speaker’ is being used? What broader issues occurred in Community X to create this situation?”)

As communities are diverse, decolonial interventions must be specific. The ideas in this essay are primarily informed by experiences of North American Indigenous communities and are offered as examples. Colonialism, however, while also realized with respect to specific places, peoples, and contexts, is manifested in Documentary Linguistics⁴ in more general ways. This is because Documentary Linguistics emerges largely from a Euroamerican colonial tradition that has guided the development of Linguistics (Errington 2008), whose scope is global but whose actors are concentrated in institutions that follow Western traditions of research. These traditions establish languages as objects to be described in scientific materials (e.g., texts, corpora, technical publications) which can serve multiple audiences, but normally are structured around colonial categories and norms of description. Examples include how languages are classified (e.g., with numerical vitality scales), named (and given ISO 639 codes), and written. Lise Dobrin and Josh Berson (2011:202) capture this pattern well in their critical analysis of language documentation, noting how “linguists’ scientific authority ... takes for granted one group’s power, derived from its association with the high-status western institution of the academy, to cast its gaze upon cultural others through the

³Terms such as *(de)colonialism*, *(de)coloniality*, *(de)colonization*, *settler colonialism*, and related concepts such as *imperialism* are used in different ways, the details of which go beyond the scope of this paper. I recommend that people engaged in language documentation focus more on the underlying ideas rather than the terms.

⁴I adopt the convention of capitalizing disciplinary names, but using lower-case to refer to the associated research.

research process, and to represent them according to its own, externally imposed analytic categories in the resulting scholarly products.”

The hegemony of academic fields is so strong that they are easily assumed to be the logical unit of analysis, and thus the default starting point from which to develop theory and research questions. For investigating (de)colonialism in Documentary Linguistics specifically, one might, for example, focus on a person whose work has had a large influence on language documentation (e.g., John Peabody Harrington), on a movement (e.g., salvage linguistics), or on a seminal publication (e.g., Himmelmann 1998). However, while potentially illuminating, a problem with this approach is that it can reproduce colonial hierarchies by elevating named academic fields over the much broader sets of lived experiences and issues that underlie language documentation needs. I thus instead draw attention to the experiences of people whose languages are the focus of recent documentation efforts.

Members of these language communities in many cases engage with Documentary Linguistics, particularly via the prototypical model whereby trained linguists from outside the community, most of whom have advanced academic credentials, work with speakers of the language under consideration to intentionally create records of it primarily for scientific purposes and secondarily to support other needs, such as language teaching. I am very familiar with this approach, but through my Miami lens, hearing the term ‘language documentation’ immediately raises other topics: The forced removal in 1846 of my direct ancestors from tribal homelands in Indiana is part of the story. The underlying intent of the Jesuit missionaries who created the first written myaamia records comes to mind as well. My ancestors’ experiences in Indian boarding schools are part of the story. My own experiences in educational institutions as a scholar who regularly has to explain basic tenets of Indigeneity, such as the fact that Native Americans still exist, are likewise part of the narrative. Although I have not personally done much direct analysis of myaamia documentation, I have heard anecdotes from others who have had trouble with it because of how myaamia is represented. Through engagement with these and similar stories, Documentary Linguistics can become more decolonial. Below, I provide a critical examination of language documentation as a colonial enterprise and offer suggestions on how to move it toward a decolonial practice.

2. Colonial approaches to language documentation Colonial approaches are not intrinsic to language documentation, as Indigenous peoples have made efforts in this area through oral traditions and cultural practices that move their languages and the associated intellectual traditions across generations. My experience, however, is that unless Indigenous community members create a systematic record of their languages in alignment with the standards of Documentary Linguistics, these efforts are not recognized as ‘language documentation’. Defining the scope of ‘documentation’ as a category is a manifestation of power. Ironically, the development of Documentary Linguistics as a named academic field, while generally beneficial for Indigenous communities, may also serve to constrain what counts.

The previous example focuses on demarcating ‘documentation’, especially with respect to its being defined separately from other activities (e.g., description) that produce similar products and are thus often similarly experienced by members of Indigenous

communities.⁵ Even more important is how ‘language’ is defined (Leonard 2017). Contemporary linguistic science privileges certain ways of defining language, particularly by structural units that can and often are described and analyzed not only separately from each other, but that are also disembodied from the people who use them, thus contradicting Indigenous values of interrelatedness as a framework for describing and interacting with the world. This trend of conceiving of languages as structurally-defined objects emerges in linguists’ analyses of Native American languages, which Joseph Errington (2008:8) observes cover “enormously different languages, [but] also resemble each other in obvious ways ... [because] each describes an object which falls under a single, common category.”

Language documentation practices can often also impose colonial norms of analyzing language in ways that misalign with the needs and values of Indigenous communities (Grenoble 2009; Hermes, Bang, & Marin 2012; Mellow 2015; Leonard 2017). ‘Dissecting’ is the word I tend to hear in critiques from Native American community members who are working with language documentation, particularly legacy documentation, and opine that some linguists’ approach to investigating language is inappropriate or offensive. This occurs, for example, when language community members encounter a linguist (or a material created by a linguist) that presents a grammatical issue as a puzzle to be solved for ‘our understanding’ (where the pronoun seems to refer to other linguists). Solving this puzzle then occurs through language data isolated from cultural contexts, and the analysis fails to acknowledge the people who claim the language, let alone to engage with what the language represents to them.

Even when all stakeholders appreciate this documentation and the people who have contributed to creating it, there is an ongoing problem of linguist-focused materials being inaccessible or otherwise misaligned with community needs. For example, a documentation corpus may contain carefully annotated texts but lack examples of conversations. The current movement in language documentation projects toward prioritizing domains that the community considers important and broadening the scope of what gets documented (e.g., by including video of conversations on diverse topics) represents a significant improvement. However, a decolonial approach calls not just for considering community needs, but rather for starting with them in conceiving of language documentation as an idea, as well as for developing the specific methods and goals of a given documentation project. Next, I present some possible interventions for accomplishing this.

3. Decolonial approaches to language documentation Were Linguistics to make a full decolonial shift, I believe that documentation would remain very important but that the norms of planning and implementing documentation projects would be driven by Indigenous research methods and protocols, which are decolonial by design and exemplified below. Due to length limitations, I focus here on recurrent themes in Indigenous research: centering details in whole systems (e.g., crafting language documentation in reference to how a language exists in its full context, and how communities want it to exist in the future), focusing on relationships and reciprocity, respecting the responsibility that comes with knowledge and its dissemination, and

⁵For the remainder of this paper, I will incorporate description, research, and products that emerge from documentation projects within my analysis of ‘language documentation’ in response to my experiences in Indigenous community settings, where these all tend to be spoken about as a single thing.

actively engaging with community needs and institutions at all stages of the research process.⁶

Using an Indigenous approach, a clear requirement for language documentation work is engagement with community definitions of ‘language’ and with community beliefs and analyses about how it functions. For instance, in my Miami community, as with many other Indigenous communities (e.g., Shaw 2001; Nelson 2002), there is a strong focus on our language’s relationship to land and an associated belief that language vitality requires community access to our lands. As such, documentation practices that fail to acknowledge the land serve to erase our connection to it, thus reinforcing the legacy of colonial violence that dispossessed Miami people of much of our land. Appropriate land acknowledgements represent a simple yet significant intervention.

Beyond an emphasis on land, also common in Indigenous community contexts is for language and peoplehood to be considered heavily intertwined (see, e.g., Clarke 1996; Meek 2010). From this point of view, representations of a language become representations of the people who claim it, and by extension also of their political sovereignty. As such, presenting the language as an object whose value lies in what it reveals for linguistic theory can reduce the people to their value for science, thus evoking the general colonial practice of exploiting the colonized population for its resources. Notably, most Indigenous community members I have spoken to about this do not mind the idea that their language’s structure will inform linguistic theory so long as the community’s ideas about language are respected. This, however, cannot easily happen, even by well-intentioned users of language documentation, unless the community’s ideas about language are present and prominent in it.

Emerging from how ‘language’ is understood are the norms of analyzing it, which also have significant implications. When documentation materials default to discrete structural units in presenting a language, the associated people can symbolically be reduced to discrete parts as well. While difficult to avoid in some situations, one useful practice when disseminating documentation is to feature larger and well-contextualized language examples, initially represented as the community would most commonly represent them (which of course first entails asking about this), and only after establishing this norm to represent them in other ways that may be necessary for specific tasks such as presenting a morphological analysis. This is especially true for language examples provided as interlinear glosses, where current disciplinary conventions allow for the initial line in a given example to be presented with the author’s morphological analysis—i.e., with hyphens, periods, and odd spacing. As argued by Wendat linguist Megan Lukaniec (2018), this is inappropriate; rather, it is important to start an example using unbroken words.

Beyond language definitions and analyses about its functions, ideas about the appropriate ways of using language must also be core to language documentation. In making this claim, I follow Jane Hill’s (2006) call for more ethnography in language documentation, where ideas about language will be emphasized on par with grammatical and lexical information. My observation has been that ethnography, though praised in its own right, is often talked about as something separate from ‘language documentation’

⁶These ideas have been developed and described by many Indigenous scholars (e.g., Wilson 2008; Kovach 2010; Smith 2012). For a short synthesis of Indigenous approaches in science, see Bang, Marin, & Medin (2018).

among linguists (except insofar as cultural patterns emerge from lexicons and texts).⁷ In all of my experiences in Indigenous communities, however, this distinction has been perceived as strange, and of course a common way for community members to define ‘language’ is with reference to culture (Leonard 2017).

Fortunately, shifts are occurring in Documentary Linguistics and in academia more widely such that colonial norms and ideas are being challenged. For instance, there are a number of projects in which community members work in successful collaborations with outside researchers who are firmly committed to supporting community needs and values in creating documentation (see, e.g., Hermes & Engman 2017; Genee & Junker 2018; and the essays in Bischoff & Jany 2018). Moreover, evidence of community engagement, support, and accessibility is normally now expected by documentation funders. A related decolonial intervention occurs with reworking legacy documentation to make it more accessible to language communities, as occurs in my Miami community.⁸ Along with this issue of accessibility, broader ethical issues are now a topic of focus, with scholarship on these issues (e.g., Grinevald 2006; Warner et al. 2007; Rice 2010, 2011) becoming standard reading for training in Documentary Linguistics. I have been especially inspired by scholarship on improving language archives by integrating community needs and insights (e.g., Linn 2014; Shepard 2016), and by the observation that community needs and insights are fundamental to a feedback loop wherein documentation, analysis, revitalization, and training work in conjunction with each other such that each is improved relative to what it would be on its own (Fitzgerald in press).

Characterizing the examples cited above is a strong awareness of how documentation is actually used (or not used) on the ground in community contexts, and how interpersonal dynamics play into the associated outcomes. From this emerges a key principle, which is that language documentation, while presumably intended to describe how a language works, can actually be prescriptive in how it is received. That is, documentation creates a baseline with prescriptive implications because it is people who create it and use it, and people have backgrounds and power relations with each other. While this is true across the board, for purposes of moving language documentation toward a decolonial practice I believe it is most useful to highlight the prototypical situation referenced earlier in which a professional linguist, who is not a community member, is the primary person who curates documentation materials and analyzes the language. When these professionals say to members of language communities things like “this is how your language’s grammar works”, their words can easily come across as fixed truths rather than what they actually are—analyses by specific people who have specific backgrounds with respect to age, gender, ethnicity, and other traits. Other practices that can yield ‘truths’, and that thus must be performed with care, include how languages are classified with respect to vitality (e.g., Leonard 2008, 2011), how speakerhood is determined and valued (e.g., Grinevald 2003; Leonard & Haynes 2010; Dobrin & Berson 2011; Muehlmann 2012; Boltokova 2017),⁹ and how languages are written with respect to orthographic choices (e.g., Romaine 2002;

⁷A reviewer notes that this split also stems from the disciplinary boundaries that emerge from colonialism and are reinforced by colonial power structures. Indeed, this is true, and breaking down disciplinary silos is part of decolonial work.

⁸Other examples of successful decolonial interventions with legacy linguist-oriented materials appear in Warner et al. (2006), Oberly et al. (2015), and Langley et al. (2018).

⁹I have observed that language consultants are too often reduced to ‘speakers’ in linguistic science, even though they are actually full people with kinship networks, occupations, responsibilities, needs, hopes, and intellectual contributions that go beyond their linguistic knowledge.

Oko 2018) along with broader issues of transcription and entextualization (see Bucholtz 2007; Riley 2009).

The effects of these decisions are heavily intertwined with the positionalities of the scholars who engage in language documentation, particularly the relationships they have with language communities and with the academy. It is thus crucial that these scholars be reflexive about how their personal backgrounds may guide how their work is perceived, and that they practice self-location (i.e., explicitly acknowledge their positionalities) in research contexts.¹⁰ The same principle applies to community members, particularly with respect to understanding how the roles that individuals are expected, allowed, encouraged, or discouraged to have in documentation projects reflect and affect their other community positions. A recurring example in my experience with small communities in the United States is that ‘elder’ gets overly linked to ‘speaker’ in documentation contexts even though this intersection of roles is not traditional, but rather a circumstance of language shift. Indeed, if a given language is known only by elders, it follows that they are going to play a pivotal role in most documentation projects; this in itself I see as sensible. The problem I have observed is that this situation too easily intersects with colonial ‘dying language’ (Leonard 2008) and ‘last speaker’ (Davis 2017) discourses that disallow younger and future generations to ever be legitimate speakers of a given language because a traditional role of elders, that of respected knowledge-bearer, has evolved through conversations about language documentation to constrain ‘real’ speakerhood. Being decolonial in language work entails calling attention to this sort of issue, which cannot be resolved unless people are aware that it might occur and thoughtful about how they advocate for, implement, and disseminate the results of documentation projects.

4. Concluding thoughts My commentary about the value of decolonial practices in language documentation is based on the following belief, whose ensuing social justice aims I have observed to be increasingly commonly proclaimed in Documentary Linguistics: Colonialism is bad, and decolonial interventions are thus appropriate. I have encountered only a few scholars who state otherwise, usually under the guise of promoting ‘objective’ science, though more common (and more troubling) are the problematic statements I have heard from language documentation practitioners who claim to support social justice but whose actions suggest otherwise. These include disparaging commentary about Indigenous groups, anecdotes about skirting tribal cultural values or research protocols, and statements that linguists are the people who truly understand language. Fortunately, this type of thinking is becoming less common, and regardless, the decolonial interventions I propose really should not be controversial, even among scientists whose goals are not focused on social justice, since these proposed practices also improve language documentation by facilitating a more complete “record of the linguistic practices and traditions of a speech community” (Himmelman 1998:166). This noted, colonialism is so engrained in the academy that abolishing it is difficult, even when doing so arguably leads to better science.

In reference to addressing this challenge, contrary to my earlier point about looking beyond academic disciplines as units of analysis, here I make a call to academic disciplines to address harmful practices through their professional structures. This can occur, for

¹⁰See Riddell et al. (2017) both for an example of researchers doing this in a publication, as well as for the authors’ excellent discussion of how self-location guides ethical research by facilitating an understanding of power differentials.

example, in a Linguistics department, which can foster a norm of including Indigenous ideas about language in introductory courses and of continuing this practice in training that is more specific to language documentation, such as field methods courses. Training in the technical aspects of documentation and linguistic analysis can be complemented by coursework in Indigenous research methods and ethnography. Professional organizations in language sciences can also advocate for decolonial practices and provide supporting guidance on the hiring, promotion, and funding of practitioners in the field. Closely related to this is the appropriate use of what is generally considered the gold standard of quality in academic work—peer review. When the language community is recognized as a core stakeholder, it follows that members of the language community will be among the reviewers of language documentation proposals and products.¹¹

I will end with my reflections on the question that I am most often asked by linguists during discussions about decolonizing language documentation: “What can I do?” Several specific decolonial practices that individuals can undertake have already been addressed above, but I intentionally leave a specific discussion about individuals’ actions for the end. I do this to illustrate yet another principle of counteracting colonialism, which is the need to shift the focus away from individual approaches in isolation to instead first center the larger social norms and institutions in which individuals operate. Stated more directly, colonialism is endemic and institutionalized; therefore, countering it in language documentation (and beyond) involves addressing colonial structures. This means that individuals must be deliberate, ideally also explicit in their research products, about counteracting colonialism, and open about where they need support. I hope in the future that decolonial forms of language documentation will be the default and thus unnecessary to call attention to, but to get to that stage we must all be thoughtful and intentional when doing language work.

¹¹Due to length limitations, I am not discussing yet another level of finalizing and disseminating this work: intellectual property and associated legal instruments, such as copyrights. I recommend Madsen (2008) for general discussion of this topic in Indigenous research, Tatsch (2004) for discussion about language as intellectual property, and Langley et al. (2018) for useful examples from a collaborative language documentation project.

References


- Bang, Megan, Ananda Marin, & Douglas Medin. 2018. If Indigenous peoples stand with the sciences, will scientists stand with us? *Dædalus: Journal of the American Academy of Arts & Sciences* 147(2). 148–159.
- Bischoff, Shannon T. & Carmen Jany (eds.). 2018. *Insights from practices in community-based research: From theory to practice around the globe*. Berlin: Mouton De Gruyter.
- Boltokova, Daria. 2017. “Will the real semi-speaker please stand up?”: Language vitality, semi-speakers, and problems of enumeration in the Canadian North. *Anthropologica* 59(1). 12–27.
- Bucholtz, Mary. 2007. Variation in transcription. *Discourse Studies* 9(6). 784–808.
- Clarke, Damon. 1996. What my Haulapai language means to me. In Gina Cantoni (ed.), *Stabilizing Indigenous languages*, 92–95. Flagstaff: Center for Excellence in Education.
- Davis, Jenny L. 2017. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. In Wesley Y. Leonard & Haley de Korne (eds.), *Language documentation and description, Vol. 14*, 37–58. London: EL Publishing.
- Dobrin, Lise M. & Josh Berson. 2011. Speakers and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 187–211. Cambridge: Cambridge University Press.
- Errington, Joseph. 2008. *Linguistics in a colonial world: A story of language, meaning, and power*. Malden, MA: Blackwell Publishing Ltd.
- Fitzgerald, Colleen M. In press. Understanding language documentation and revitalization as a feedback loop. In Stephen Fafulas (ed.), *Amazonian Spanish: Language contact and evolution*. Amsterdam: John Benjamins.
- Fitzgerald, Colleen M. & Mary S. Linn. 2013. Training communities, training graduate students: The 2012 Oklahoma Breath of Life workshop. *Language Documentation & Conservation* 7. 185–206.
- Genee, Inge & Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary Project: Creating integrated web resources for language documentation and revitalization. *Language Documentation & Conservation* 12. 274–314.
- Grenoble, Lenore A. 2009. Linguistic cages and the limits of linguists. In Jon Reyhner & Louise Lockard (eds.), *Indigenous language revitalization: Encouragement, guidance & lessons learned*, 61–69. Flagstaff: Northern Arizona University.
- Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In Peter K. Austin (ed.), *Language Documentation and Description, Volume 1*, 52–72. London: SOAS.
- Grinevald, Colette. 2006. Worrying about ethics and wondering about “informed consent”: Fieldwork from an Americanist perspective. In Anju Saxena & Lars Borin (eds.), *Lesser-known languages of South Asia: Status and policies, case studies and applications of information technology*, 339–370. Berlin: Mouton de Gruyter.
- Hermes, Mary, Megan Bang, & Ananda Marin. 2012. Designing Indigenous language revitalization. *Harvard Educational Review* 82(3). 381–402.
- Hermes, Mary & Mel M. Engman. 2017. Resounding the clarion call: Indigenous language learners and documentation. In Wesley Y. Leonard & Haley de Korne (eds.), *Language Documentation and Description, Vol 14*, 59–87. London: EL Publishing.
- Hill, Jane H. 2006. The ethnography of language and language documentation. In Jost Gippert, Nikolaus P. Himmelmann, & Ulrike Mosel (eds.), *Essentials of language documentation*, 113–128. Berlin: Mouton de Gruyter.

- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195.
- Hinton, Leanne. 2001. The use of linguistic archives in language revitalization: The Native California language-restoration workshop. In Leanne Hinton & Ken Hale (eds.), *The green book of language revitalization in practice*, 419–423. San Diego: Academic Press.
- Kovach, Margaret. 2009. *Indigenous methodologies: Characteristics, conversations, and contexts*. Toronto: University of Toronto Press.
- Langley, Bertney, Linda Langley, Jack B. Martin, & Stephanie Hasselbacher. 2018. The Koasati Language Project: A collaborative, community-based language documentation and revitalization model. In Shannon T. Bischoff & Carmen Jany (eds.), *Insights from practices in community-based research: From theory to practice around the globe*, 132–150. Berlin: Mouton de Gruyter.
- Leonard, Wesley Y. 2008. When is an “extinct language” not extinct?: Miami, a formerly sleeping language. In Kendall A. King, Natalie Schilling-Estes, Lyn Fogle, Jia Jackie Lou, & Barbara Soukup (eds.), *Sustaining linguistic diversity: Endangered and minority languages and language varieties*, 23–33. Washington: Georgetown University Press.
- Leonard, Wesley Y. 2011. Challenging “extinction” through modern Miami language practices. *American Indian Culture and Research Journal* 35(2). 135–160.
- Leonard, Wesley Y. 2012. Framing language reclamation programmes for everybody’s empowerment. *Gender & Language* 6(2). 339–367.
- Leonard, Wesley Y. 2017. Producing language reclamation by decolonising ‘language’. In Wesley Y. Leonard & Haley de Korne (eds.), *Language Documentation and Description Vol 14*, 15–36. London: EL Publishing.
- Leonard, Wesley Y. & Erin Haynes. 2010. Making “collaboration” collaborative: An examination of perspectives that frame field research. *Language Documentation & Conservation* 4. 268–293.
- Linn, Mary. 2014. Living archives: A community-based language archive model. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description Vol 12*, 53–67. London: SOAS.
- Lukaniec, Megan. 2018. Supporting Native and Indigenous linguists in academia. Paper presented at the 92nd Annual Meeting of the Linguistic Society of America, 4–7 January 2018.
- Madsen, Kenneth D. 2008. Indigenous research, publishing, and intellectual property. *American Indian Culture and Research Journal* 32(3). 89–105.
- Meek, Barbra A. 2010. *We are our language: An ethnography of language revitalization in a Northern Athabaskan community*. Tucson: University of Arizona Press.
- Mellow, J. Dean. 2015. Decolonizing Western science, research, and education: Valuing linguistic diversity. In Jon Reyhner, Joseph Martin, Louise Lockard, & Willard Sakiestewa Gilbert (eds.), *Honoring our elders: Culturally appropriate approaches for teaching Indigenous students*, 45–60. Flagstaff: Northern Arizona University.
- Muehlmann, Shaylih. 2012. Von Humboldt’s parrot and the countdown of last speakers in the Colorado Delta. *Language & Communication* 32. 160–168.
- Nelson, Melissa. 2002. Introduction: Indigenous language revitalization. *ReVision* 25(2). 3–4.
- Oberly, Stacey, Dedra White, Arlene Millich, Mary Inez Cloud, Lillian Seibel, Crystal Ivey, & Lorelei Cloud. 2015. Southern Ute grassroots language revitalization. *Language Documentation & Conservation* 9. 324–343.

- Oko, Christina Willis. 2018. Orthography development for Darma (The case that wasn't). *Language Documentation & Conservation* 12. 15–46.
- Rice, Keren. 2010. The linguist's responsibilities to the community of speakers: Community-based research. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language documentation: Practice and values*, 25–36. Amsterdam: John Benjamins.
- Rice, Keren. 2011. Ethics in fieldwork. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 407–429. Oxford: Oxford University Press.
- Riddell, Julia K., Angela Salamanca, Debra J. Pepler, Shelley Cardinal, & Onowa McIvor. 2017. Laying the groundwork: A practical guide for ethical research with Indigenous communities. *The International Indigenous Policy Journal* 8(2).
- Riley, Kathleen C. 2009. Who made the soup? Socializing the researcher and shaping her data. *Language & Communication* 29. 254–270.
- Romaine, Suzanne. 2002. Signs of identity, signs of discord: Glottal goofs and the green grocer's glottal in debates on Hawaiian orthography. *Journal of Linguistic Anthropology* 12(2). 189–224.
- Sammons, Olivia N. & Wesley Y. Leonard. 2015. Breathing new life into Algonquian languages: Lessons from the Breath of Life Archival Institute for Indigenous Languages. In J. Randolph Valentine & Monica Macaulay (eds.), *Papers of the 43rd Annual Algonquian Conference*, 207–224. Albany: SUNY Press.
- Shaw, Patricia A. 2001. Language and identity, language and the land. *BC Studies: The British Columbian Quarterly* 131. 39–55.
- Shepard, Michael Alvarez. 2016. The value-added language archive: Increasing cultural compatibility for Native American communities. *Language Documentation & Conservation* 10. 458–479.
- Smith, Linda Tuhiwai. 2012. *Decolonizing methodologies: Research and Indigenous peoples*. 2nd edn. New York: Zed Books.
- Tatsch, Sheri. 2004. Language revitalization in Native North America—issues of intellectual property rights and intellectual sovereignty. *Collegium Antropologicum* 28(suppl. 1). 257–262.
- Warner, Natasha, Lynnika Butler, & Quirina Luna-Costillas. 2006. Making a dictionary for community use in language revitalization: The case of Mutsun. *International Journal of Lexicography* 19(3). 257–285.
- Warner, Natasha, Quirina Luna, & Lynnika Butler. 2007. Ethics and revitalization of dormant languages: The Mutsun language. *Language Documentation & Conservation* 1(1). 58–76.
- Wilson, Shawn. 2008. *Research is ceremony: Indigenous research methods*. Black Point, Nova Scotia: Fernwood Publishing Company.

Wesley Y. Leonard

wesley.leonard@ucr.edu

 orcid.org/0000-0001-8792-4414

Reflections on public awareness

Mary S. Linn

Smithsonian Center for Folklife and Cultural Heritage

In this reflection, I repeat Michael Krauss’s 1992 call for linguists of all kinds to be active in creating public awareness of language endangerment, and more importantly at this stage, in motivating global attitudinal changes in support of language diversity. I purposely do not distinguish between academic and non-academic, community and non-community linguists, requiring that we all participate in this call. I distinguish different target publics, namely the endangered or minoritized language community public and the majority language public in terms of message and response. I then briefly outline past and present efforts in varying media that are part of creating awareness and action on a global scale. I focus on integration of media and message, stressing that we must be able to provide a positive vision of a linguistically diverse world and a means for the general public, especially youth, to participate in its creation.

1. Introduction In Hale et al. 1992, after presenting the eye-opening estimate that without intervention up to 90% of the world’s language would disappear by the end of the 21st century, Michael Krauss asked, “What are we linguists doing to prepare for this or to prevent this catastrophic destruction of the linguistic world?” (p. 7). One way that linguists could prepare, and the way most suited to linguists, was to document endangered languages and to document in a way that benefitted both community endeavors and linguistic science. Krauss’s call to prepare thus signaled a slow but steady restoration of field methods, or *practicum*, courses in linguistics departments. Himmelmann 1998 greatly impacted the legitimacy of descriptive linguistics. By disentangling and systematizing the terms *documentation* and *description*, he escalated the growth from what a few dedicated linguists were doing into a distinct subfield of linguistics, now called documentary and descriptive linguistics. Himmelmann’s stated motivation for systematizing the field was his concern for endangered languages (p. 161). As the subfield has grown, so has public awareness of language endangerment and loss.

The second part of Krauss’s question, what we are doing to *prevent* the catastrophic loss of languages, is much harder. Many academic linguists come to the field of documentation and description through a love for the languages themselves. Most

come to language revitalization or reclamation by earnest commitment to and with an endangered language community and community members they work with. However, prevention is inherently tied to our ability to interact not with language data or with a community of speakers and advocates that share similar if not the same goals, but with the general public.

Unfortunately, the general public still does not usually share our love of language itself, and in many cases feels little or no sympathy towards language loss and the people who are affected. The general public creates donors, and we do need money to implement language revitalization. The general public creates local, national, and international politicians who create policy, and we need supporting policy to affect language reclamation. The general public, most importantly, can create a climate for language revitalization and diversity to flourish. If we are successful in creating new generations of speakers and renewing healthy language communities but the speakers of majority languages are not accepting of language diversity, then small languages will always struggle to maintain a foothold. Without a wider climate of support, we are at best creating at-risk enclaves, and at worst creating margins of future conflict.

A quick view of recent comments posted on-line after endangered language-related articles, even in the more educated outlets such as the BBC and the New York Times, reveal common attitudes towards endangered languages and their speakers. Leaving outright xenophobic and racist comments aside (and there is a lot of it), the negative attitudes fall into three main groups: 1) Those who still think, despite overwhelming evidence to the contrary, that a common language creates less conflict, 2) those who cling to the discredited views of Social Darwinism and believe that language loss is simply the way of the world, and that languages will die off to 'stronger' languages, cultures and economies, so there is nothing to be done about it. This is often expressed by 'that's progress' or 'just get over it' or 'the will of God,' and 3) those who see and perhaps understand the issues, but are apathetic because they do not see how language loss impacts them.

Prevention begins with awareness, but ultimately, we must change the attitudes and behavior of majority language speakers, especially those in predominately monolingual societies and spaces such as the internet. But how do we linguists mobilize a global society when the underlying causes of language endangerment are colonialism, poverty, xenophobia, and racism? How do we combat the economics of globalization that push individuals and communities to adopt majority languages? Linguists are not trained to do this, nor do many of us feel comfortable in the role of spokespersons for a cause. Yet, we have to. We cannot live in this time and be linguists and do nothing. Even if it feels very small, if each one of us does something, it will add up.

In this reflection, when I implore action or give credit to action with the words *we* or *linguists*, I am not speaking narrowly of academically trained career linguists. By linguists, I mean language practitioners (endangered language community members working on language outside of academia), community linguists working in academia, non-community linguists working with communities (revitalization/reclamation) and those working with endangered languages (documentation and description), and even non-community linguists who do not work with communities or endangered languages. We all bear responsibility for educating the general public to the causes and effects of language loss, and the tireless efforts and milestones achieved in reversing language shift. That being said, since language practitioners and community linguists are often overworked and over stretched in their own communities, academic linguists working in majority-language settings should take up the liaison role of raising public awareness.

By *general public*, I mean speakers of majority languages whose lives are not normally touched by causes or effects of language loss but can be supportive of language diversity and cultural plurality around them and can become agents in ending discriminatory practices and preferential treatments.

My linguistic career has led me to two very public positions, first as the Curator of Native American Languages at the Sam Noble Museum in Norman, Oklahoma, and second as Curator of Cultural and Linguistic Revitalization at the Smithsonian Center for Folklife and Cultural Heritage in Washington, DC. Museums are inherently institutions of public education. More and more museums have permanent and temporary galleries dedicated to endangered languages. Most are in collaboration with communities, many include youth participation, and some have accompanying education materials. For a few examples, see The Smithsonian Folklife Festival 2013 *One World Many Voices* program,¹ Royal BC Museum *Our Living Languages*,² *First Peoples* at the Melbourne Museum,³ and The Canadian Language Museum includes a helpful map of language museums around the world.⁴ The yearly Oklahoma Native American Youth Language Fair at the Sam Noble Museum includes the mission of presenting living languages. Lena Herzog's *Last Whispers: Oratorio for Vanishing Voices, Collapsing Universes, and a Fallen Tree* is an immersive installation that premiered at the British Museum in 2016 and is traveling to other major museums.⁵ Planet Word, a museum dedicated to language is opening in 2019 in Washington, DC, and will have sections about endangered languages.⁶

As an employee of publicly funded museums, I have taken on a role to educate the public about the causes and effects of language shift, the amazing strides achieved in communities and schools, and the positive steps that all people can play in reversing language shift. I have learned, and am still learning, how to present linguistic and cultural issues to the public. I hope to share some of my lessons and thoughts in this reflection.

2. From awareness to mobilization Twenty years ago, the wake-up call was mainly aimed at linguists, and 'the public' was mainly endangered and minoritized language communities. As with the linguistic community, many communities did not fully realize the extent of language shift happening at home, let alone what their communities were struggling with was shared by minoritized and small language communities around the world. Communities did not understand the ramifications of generations of school-aged children not speaking the language in the home, or of having only middle-aged or older first language speakers. While it seems nearly impossible today, there are still endangered language communities that are not aware of their language loss. These are mainly in less economically privileged areas of the world, where speakers of small languages are focused on day-to-day living, or are pulled by economic pressures into urban areas for jobs to survive. And in some instances, active language revitalization may be politically dangerous, and so bringing awareness on the issues can be harmful as well.

¹<https://festival.si.edu/2013/one-world-many-voices/smithsonian>, and many of the past Smithsonian Folklife Festival programs emphasize language and central to heritage transmission and include language teachers and lessons.

²<https://royalbcmuseum.bc.ca/visit/exhibitions/our-living-languages-first-peoples-voices-bc>

³<https://museumsvictoria.com.au/website/bunjilaka/visiting/first-peoples/>

⁴<http://www.languagemuseum.ca/language-museum-map>

⁵<http://www.lastwhispers.org/>

⁶<https://www.planetwordmuseum.org/>

Yet, when awareness comes, most communities jump swiftly from awareness to mobilization. The desire to act very quickly generally outpaces organization, funding, and training. The mid-1990's and early 21st century saw a dramatic shift from how to get youth motivated to being able to providing enough teachers and resources to keep up with youth demand. Youth are the driving force and often the practitioners of language reclamation efforts today. In many parts of the world, young adults and youth have grown up knowing about language loss and revitalization. This awareness is part of their everyday consciousness growing up, and we have yet to know how fully this will play out in the next decades of reversing language shift.

One of the first community responses was to get training in language documentation, description, language teaching methodologies, and literature development. Semi-formal and informal training institutes in linguistics and revitalization approaches and methodologies (see Fitzgerald in this volume), with more and more higher degree programs available in language revitalization and in Indigenous languages. Early trainees helped spread the word in their own communities, and in addition to creating language practitioners, an active network of Indigenous language advocates sprang up. *Language advocacy* is a recognized need and role in many endangered language communities. Today, the Resource Network for Linguistic Diversity in Australia and Collaborative Language Research Institute in North America have regular workshops in language advocacy. These workshops instruct not just how to convey the message and efforts within communities, but how to actively engage the wider public for support, funding, and policy. These are skills not yet taught in linguistics or anthropology departments, or cross-listed with other departments as acceptable core credit for linguistics degrees.

Not enough credit is given to Indigenous language speakers in early public awareness of language endangerment. While community-wide awareness in endangered language communities may have come in part via outside linguists or linguist-driven media coverage, the linguists' knowledge came from concerned community members with whom they worked. Linguists since Boas's time have been aware of language shift, but it was not until the cultural re-awakenings of the late 1960's (coupled in the US with opening of the bilingual education and in Europe with political shifts in the 1970s and 80s) that speakers and consequently their linguists began taking serious action against language shift. So, in endangered language communities, awareness has come from within through speakers and speaker communities, who were also often the first community linguists, the early trainees at language institutes, and from outside through linguists and their efforts to publicize widely. More crucially, awareness (and effective approaches) in endangered language communities comes laterally across the many historical and modern connections that bind Indigenous and autochthonous people together locally and globally.

3. The message itself The general public must be made aware of the causes and scope of language shift: an overall picture of linguistic diversity and the overwhelming and systematic loss of small and minoritized languages to majority languages and dominating political, economic, and social structures. We have done a fairly good job of this in a relatively short period of time. The largest surge of public awareness came in 1990-2010. In 1991, the Linguistic Society of America organized a symposium on language endangerment, leading to 1992 publication in *Language* of Hale et al., including Krauss's clarion call mentioned above. In 1992, the 15th International Congress of Linguists meeting in Quebec raised awareness of language shift and loss to the international linguistics community, and their statement of urgency directly influenced the UNESCO

General Assembly to create the more outwardly focused *Red Book on Endangered Languages* (1993). The *Red Book* went online in 2009 and is now the digital *UNESCO Atlas of the World's Languages in Danger*.⁷ These decades produced supportive language policy (always an opportunity for public engagement), non-profit agencies (with dedicated public awareness missions), funding agencies (also good opportunities for press releases), several popular press books,⁸ and Indigenous language institutes and conferences. Most of these are still impactful today and are legacy to many of our current efforts.

Crystal (2011) points out that it has taken the biological world much longer to educate the world about environmental loss and endangered species. For example, between 2007-2013 National Geographic's Enduring Voices Project, with Gregory Anderson, K. David Harrison, and Chris Rainer (photographer), exponentially elevated language endangerment as a global issue, a little over 10 years from Himmelmann. The Audubon Society, on the other hand, has been working for over 100 years to make us aware of the decline in bird species and populations. The upcoming UNESCO Year of Indigenous Language 2019 is a good opportunity for a next large push in public awareness.

There is an intricate balance between raising awareness and producing negativity, blame, and guilt for language communities. While we refuse the rhetoric of *dead* or *extinct* languages to *sleeping* or languages when there are communities who are renewing languages with no current first language fluent speakers, we need to communicate the scope of the crises. And while we refuse the rhetoric that a culture will cease to be when the last speaker dies, we need to communicate the impact of language shift on the community.⁹ If we must pull the public in (the scare), then we must keep them going with the dream, and the hope through the reawakening and language renaissances taking place in and across Indigenous communities. Indigenous and minoritized groups have this: Most can and do envision their community with their languages maintained, or renewed, and this entails for them a healthier, more educated (in their own definitions of educated) citizens. Indeed, language revitalization is a movement in which communities (re)define their groupness and frame other social and political rights (Costa 2017). We cannot just produce fear or guilt on the part of the general public as well. Our task is to build a positive vision for a linguistically diverse future that positively impacts speakers of majority languages.

The media is in continual shock mode, creating an emotional overload to the point that people cannot or do not react. Climate change and its induced natural disasters, poverty and inequalities in justice and education, health crises in Alzheimer's Disease, diabetes, substance abuse and suicide, and nationalistic and imperialistic behavior that spawns continual localized wars and threatens new global wars. A crisis in language just becomes part of the noise. We cannot compete with these if we continue to see language as separate from these other issues. We must admit that most people do not get into language by itself; it is too esoteric, too hard, too remote. Without a positive vision, the general public will continue to be unmoved, or feel helpless at best, in seeing traditional or minoritized cultures relent to urbanization and globalization.

All these major world problems are interrelated and have the same causes as those which create language shift. I firmly believe that recreating and sustaining healthy language communities is part of the solution for all of them. Language renewal is

⁷<http://www.unesco.org/languages-atlas/>

⁸See Hagège 2000, Romaine & Nettle 2001, Crystal 2001, Abley 2003, and Harrison 2007.

⁹For fuller discussions on the rhetoric of language revitalization and its impact, see Hill (2002) and the responses, Perley (2012), Heller & Duchêne (2017), and De Korne & Leonard (2017), among others.

inextricably related inequalities and poverty, to health and suicide rates, and to education. In renewing healthy language communities, we are creating combating prejudices and inequalities, we are creating healthier living, eating, and communities, and we are taking back control of the education of our youth.¹⁰ These problems are so overwhelming and out of people's daily control that language renewal suddenly seems like a part of the solution that people can grasp.

We give them a succinct role, just like the environmental movement has provided consciousness in campaigns like "Reduce, Recycle, Reuse" that promote recycling, taking shorter showers, replacing older light bulbs, planting green spaces and so on. People can understand their individual role in a global problem. Not everyone participates, but enough do that it makes a difference. And importantly, the environmental movement has mobilized younger generations. In the US today, the dividing lines between the two political parties are further apart than most times in our history. However, the youth cross party lines today to support environmental initiatives and confront global warming.

What is the role of the general public in endangered languages? In an increasingly less empathetic world, it is a beginning just to listen to the communities. Be supportive local languages in the schools, and all community revitalization and reclamation endeavors. In monolingual societies, work towards bilingual adults, if not a multilingual society by supporting even majority second language acquisition in primary schools. Easier than arguing for tolerance, we can argue brain health and better pay scales for this, and by raising bilingual adults, there will be much less fear of smaller languages and of others.¹¹ Citing Crystal, "This is not such a great effort as it may appear, compared with the efforts that go into much more dubious enterprises. And let us not forget that the costs of war are always greater than those of peace."¹²

4. A new hope for systemic change We are in a new era of participatory culture through the internet and social media, with consequences for public awareness and acceptance of linguistic diversity. The efforts may seem more diffuse than in the 1990s. However, through video streaming, podcasts, Tedx Talks, blogs and vlogs, and memes (to name a few), we have the capacity to move the message more quickly and to involve more people. In particular, the internet combines the message with the arts, and the arts have the power to move people, to make them feel empathy with the subject. The power of the internet in the diffusion of Indigenous hip hop, rap, and slam poetry as a vehicle for youth expression inspiring youth all over the globe cannot be underestimated.

Film and video are probably the strongest artistic medium to inform and motivate to action the largest number of people. Because of film's popularity throughout the world, it is increasingly cheaper to produce high quality film and it is increasingly easier to distribute widely through the internet. Smart phones and cheaper hand-held cameras are providing a hitherto unknown level of Indigenous youth voices in film. A few documentaries have tackled the subject of language shift, culture, identity,

¹⁰For evidence of health and educational benefits in language reclamation, see Whalen et al. (2016), Child Language Research and Revitalization Working Group (2017), Taff et al. (2017), and Fitzgerald (2017), among others.

¹¹An effective example of the social media messaging can be found in the America Versus video series on Facebook. the May 14 installment entitled America Vs Language effectively presents these arguments by comparing monolingual America with other countries.

¹²In his keynote address to the Barcelona Congress in May 2004, David Crystal gave ten specific measures for mobilizing society as a whole (cited in Mari 2008: 91). I encourage everyone to read them in full. Many of his proposals are underway.

documentation and revitalization efforts worldwide.¹³ Other documentaries portraits of specific language communities, their histories and issues of historical trauma and healing connected to language loss and revitalization.¹⁴ Exceeding both of these genres is Indigenous film. Many go straight to downloading via the internet, but more and more they are screened at internationally recognized film festivals and at least 87 Indigenous film festivals worldwide (Cordova 2015). Festivals and accompanying video releases give an Indigenous voice to public discussion around the issues surrounding cultural loss, trauma and renewal, and approaches for sustainability.¹⁵ The interconnectedness and the beauty of these stories, along with the languages, seeps into the consciousness of generations, as does comfortableness with 'the other.' Films move us to care about the people, the places, the cultures.

Working towards long-term, systemic change in public attitudes, the best place to start is with the youth. We need to have our message reach young people while they are forming their views of the world around them. A few linguists actively engage with local schools to teach about language endangerment, but more of us need to connect with teachers and youth in this way, even if it is one guest lecture to a class or school club. Some organizations have teaching materials or information for teachers on their websites.¹⁶ To be broadly impactful, our field needs to develop many resources for teachers to use or able to be modified for a variety of class types (formal and informal settings) for a full range of learning levels.

I have hope in change through youth. I recently looked through a yearbook from a high school in Wichita, Kansas, the rural, conservative, monolingual heartland of the US. This was the same high school that I graduated from nearly forty years ago. When I was there, I know we had students who spoke Spanish at home, and we had the first speakers of Southeast Asian languages coming into the school attending mainly ESL classes, but I never heard a language other than English spoken outside of a foreign language class. The 2018 student-run yearbook contains 13 testimonies about what the high school means to them. These testimonials are in 13 languages other than English, and then translated into English.¹⁷ The testimonials frame the yearbook, with half at the beginning and half at the end. They are a clear celebration of their diversity, and an unstated, natural endorsement of a multilingual world.

¹³See films such as *The Linguists* (Kramer, Miller & Newberger 2008) and *Language Matters with Bob Holman* (Grubin 2014).

¹⁴*We Still Live Here – Às Nutayuneân* (Makepeace 2011), and *Keep Talking* (Weinberg 2017) are two excellent examples of this genre.

¹⁵The Smithsonian Mother Tongue Film Festival (MTFF) in Washington, DC, begins every year on UNESCO International Mother Language Day on February 21. While focusing on Indigenous film, MTFF is dedicated to films in endangered and minoritized languages or films about language endangerment and renewal.

¹⁶The Stolen Generation, an educational website sponsored by the Australian government, has in-depth teaching resources, including curriculum and sequences learning modules. Terralingua has a biocultural education initiative (<http://terralingua.org/our-work/bcd-education/>) The 2015 interactive map Native Land that allows people to overlay current Indigenous language boundaries with historical treaties and traditional lands of former British colonies the site includes information for teachers (<https://native-land.ca/>). The Endangered Language Project has recently hired staff to create learning modules for teachers to better use the site and to incorporate lessons on endangered languages into their curriculum (<http://www.endangeredlanguages.com>).

¹⁷In their words, these languages are Vietnamese, Spanish, Kinyarwanda, Bengali, Uganda, Algerian French, Congo Swahili, Swahili, Cambodian, Bangla, Arabic, Turkish.


References

- Child Language Research and Revitalization Working Group. 2017. *Language documentation, revitalization, and reclamation: Supporting young learners and their communities*. Waltham, MA: EDC.
- Cordova, Amalia. 2015. *Nomadic/Sporadic: The pathways of circulation of Indigenous video in Latin America*. PhD Dissertation, New York University.
- Costa, James. 2017. *Revitalising language in Provence: A critical approach*. New York: Wiley.
- Crystal, David. 2001. *Language death*. Cambridge University Press.
- Crystal, David. 2011. Language diversity, endangerment, and public awareness. Keynote address for The British Academy, London, 23 February. (<https://soundcloud.com/britishacademy/language-diversity-endangerment-and-public-awareness>) (Accessed 2018-08-02)
- De Korne, Haley & Leonard, Wesley Y. 2017. Reclaiming languages: Contesting and decolonising 'language endangerment' from the ground up. In Wesley Y. Leonard & Haley De Korne (eds.), *Language Documentation and Description Vol 14*, 5–14. London: EL Publishing.
- Leonard, Wesley Y. 2017. Producing language reclamation by decolonizing 'language'. In Wesley Y. Leonard & Haley De Korne (eds.), *Language Documentation and Description Vol 14*, 15–36. London: EL Publishing.
- Fitzgerald, Colleen. 2017. Understanding language vitality and reclamation as resilience: A framework for language endangerment and 'loss' (Commentary on Mufwene). *Language* 93. e280–e297.
- Taff, Alice, Melvatha Chee, Jaeci, Hall, Millie Yéi, Dulitseen Hall, Kawenniyóhstha Nicole Martin & Annie Johnston. 2017. Indigenous language use impacts wellness. In Kenneth Rehg & Lyle Campbell, *The Oxford handbook of endangered languages*, 862–883. New York: Oxford University Press.
- Grubin, David. 2014. *Language Matters with Bob Holman*. USA: David Grubin Productions.
- Hale, Kenneth L., Colette Craig, Nora England, LaVerne Jeanne, Michael Krauss, Lucille Watahomigie & Akira Yamamoto. 1992. Endangered languages. *Language* 68. 1–42.
- de Hagege, Claude. 2000. *Halte à la mort des langues*. Paris: Odile Jacob.
- Heller, Monica & Alexandre Duchêne. 2007. Discourses of endangerment: Sociolinguistics, globalization and social order. In Alexandre Duchêne & Monica Heller (eds.), *Discourses of endangerment: ideology and interest in the defence of languages*, 1–13. London: Continuum.
- Hill, Jane H. 2002. 'Expert rhetorics' in advocacy for endangered languages: Who is listening and what do they hear? *Journal of Linguistic Anthropology* 12. 119–133.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Kramer, Seth, Daniel A. Miller & Jeremy Newberger. 2008. *The linguists*. USA: Ironbound Films.
- Makepeace, Ann. 2011. *We still live here – Âs nutayuneân*. USA: Makepeace Productions.

- Marí, Isador. 2004. Globalisation and linguistics rights: towards a universal framework of linguistic sustainability. Inaugural lecture at the XIV Seminario de la Enseñanza de Lenguas Extranjeras, 'La diversidad lingüística en el contexto de la globalización', an academic function held in parallel to the xviii Feria Internacional del Libro de Guadalajara, 1–2 December 2004. (https://11u11.cat/IMAGES_175/transfer01-foc03.pdf) (Accessed 2018-08-02)
- Perley, Bernard C. 2012. Zombie linguistics: experts, endangered languages and the curse of undead voices. *Anthropological Form* 22. 133–149.
- Weinberg, Karen Lynn. 2017. *Keep talking*. USA: Kartemquin Films.
- Whalen, Douglas H., Margaret Moss & Daryl Baldwin. 2016. Healing through language: positive physical health effects of indigenous language use. *F1000Research* 2016, 5.852 (doi:10.12688/f1000research.8656.1)

Mary S. Linn

linnm@si.edu

 orcid.org/0000-0002-9739-2690

Key Issues in Language Documentation



Interdisciplinary research in language documentation

Susan D. Penfield
University of Montana
University of Arizona

This paper explores the parameters of interdisciplinary work in language documentation. Citing the strong call for the involvement of disciplines, other than linguistics, beginning with Himmelmann, to the present trajectories for language documentation research, the author claims that more attention is needed to the enactment of interdisciplinary work from project conception to the follow-through in terms of where to disseminate outcome.

1. The Past I was lucky. My early fieldwork experience, as a first-year graduate student, was guided by an oral history project under which I was supposed to collect some ‘native language.’ This allowed me to get some experience with data collection of various types and to see a wider range of texts than if the task had been straight descriptive linguistics. This experience led to a career as a linguistic anthropologist, where the concept of interdisciplinary work has never been far from my thoughts. Linguistics, when I was a graduate student in the late 1960s, was still considered a sub-field of anthropology in the United States and a required subject for anthropology majors. As such, it was strongly rooted in descriptive linguistics, as put forth by Bloomfield (1934) and the Structuralist era that followed. Linguistic fieldwork, in the Boasian tradition of creating grammars, texts and dictionaries, was still the norm. As Evans points out, this trilogy, while useful, “will not supply all the questions that future linguists and community members will want to ask” (2010:223). Against this backdrop, students of linguistic anthropology proceeded to describe languages with a focus on the transcription and analysis of collected primary data from unwritten languages. These same students also studied language use and practices, the more anthropological side of the equation, quite independently of any linguistic description. While the disciplines of linguistics and anthropology were generally seen as closely related, the nature of that relationship did not seem clearly defined, at least to me. One could do a lot of descriptive linguistics without much, if any, appeal to the more anthropological concerns of how people might actually use the language in given situations and for different purposes.

There evolved parallel areas of research—a strong tradition of careful language descriptions sometimes in concert with, but not blended with, anthropological perspectives on language.¹ I think that to most of us, who had been trained in descriptive linguistics, the entrance of language documentation was truly eye-opening and welcome.

What I find most interesting about Himmelmann’s early writings is that, from the beginning, language documentation mandates an interdisciplinary approach. The door to that is heavily messaged in his article, “Documentary and Descriptive Linguistics” (1998), beginning with this point that, “Language descriptions are, in general, useful only to grammatically oriented and comparative linguists. Collections of primary data have at least the potential of being of use to a larger group of interested parties. These include the speech community itself, which might be interested in a record of its linguistic practices and traditions” (1998:63). He adds that a set of primary data may be of interest to various other (sub-)disciplines, including sociolinguistics, anthropology, discourse analysis, oral history, etc. This, of course, presupposes that the data set contains data and information amenable to the research methodologies of these disciplines (1998:163). This last point has become increasingly important as interdisciplinary work in language documentation has evolved and is even more in focus for the future, as I will comment on below. As well, the mention of data being of interest to the speech community itself, while not revolutionary, was certainly overdue, and spoke to the need to ‘give back’ and share outcomes. How different my early linguistic fieldwork experience might have been if I had been trained to think more broadly and how interesting it is, at this juncture, to think about how language documentation so clearly, and so deeply, embraces interdisciplinary work.

The mandate to include multiple disciplines is clearly spelled out in the statement that a language documentation should include a “comprehensive record of the linguistic practices characteristic of a given speech community” (Himmelmann 1998:166). While it is possible to consider language structure and language use within the field of linguistics, the notion of documenting ‘linguistic practices’ broadly goes beyond the strict documentation of linguistic forms and aims to understand that the documentation of language must be broad enough to consider language structure and use *related* to changing topics, events, places, individuals and more. Even further, Himmelmann writes that “a language documentation aims at the record of the linguistic practices and traditions of a speech community” (1998:166). Specifically, he adds that the “makeup and contents of a language documentation are determined and influenced by a broad variety of language related (sub-disciplines) including: sociological and anthropological approaches to language ... ‘hardcore’ linguistics (theoretical, comparative, descriptive); discourse analysis, spoken language research, rhetoric; language acquisition; phonetics; ethics, language rights, and language planning; field methods; oral literature and oral history; corpus linguistics; educational linguistics” (1998:167). The list has grown since then. Himmelmann adds, “The major theoretical challenge for documentary linguistics is the task of synthesizing a coherent framework for language documentation from all of these disciplines” (1998:167). And, it is this final point that has presented the greatest challenge to interdisciplinary studies within the framework of language documentation.

Thus, the trajectory from descriptive linguistics to language documentation, at its core, has taken us from a single focused exercise in describing any given language, to a multi-faceted effort, inclusive of the description but moving beyond it, casting a wider net that encompasses language-related practices and the various disciplines that might interface

¹For further insights on this see Epps et al. (2017).

with them. The strength and opportunity provided by expanding language documentation projects into other disciplines was obvious from the beginning and, as noted above, so were the anticipated challenges.

2. The Present I will take ‘the present’ to be roughly from 1998, the time of Himmelmann’s seminal article defining language documentation, to the time of this writing in 2018. Language documentation has advanced at a steady pace to a fully recognized, stand alone, new field of study in a relatively short time period. The interest in interdisciplinary perspectives has intensified and somewhat changed over this period of time. We can simply reflect on this growth by looking at the papers written for two significant publications that took place during this period: The first is *The Essentials of Language Documentation* (2006), edited by Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel; the second, *The Oxford Handbook of Linguistic Fieldwork* (2012), edited by Nicholas Thieberger.

In the 2006 volume, the papers were still tightly tied to the mechanics of performing language documentation, for the most part. Understandably, the advent of advanced technology marks one of the biggest shifts distinguishing language documentation. More sophisticated technology and the ability to digitally store, analyze and managed bulks of data continues to be one of the interesting challenges of the field. Early on, there was an intense focus on what technology brought to the table and on how to best use it. In 2006, the field, as a whole, was still working through the challenges of new recording possibilities, new data management issues, and new spins on what technology brought to descriptive work, such as E-MELD.² Papers such as, “Data and language documentation” (Austin), “Documenting lexical knowledge” (Haviland), “Linguistic annotation” (Schultze-Berndt), “Archiving challenges” (Trilsbeek & Wittenburg) were indicative of this. There were also indicators of the interdisciplinary work and considerations of speech communities evident in this early volume, note the papers about “Ethics and practicalities of cooperative fieldwork and analysis” (Dwyer), “Ethnography in language documentation (Franchetto), and “Fieldwork and community language work” (Mosel). At the same time, those who were engaged in fieldwork were increasingly experimenting with interdisciplinary applications through the lens of language documentation. That trend grew markedly by 2012.

In Thieberger’s 2012 book, the papers still address issues in data recording and management, but many reflect broader trends in the interpretation of language documentation as a site for interdisciplinary studies. It is almost as if, once the field became more comfortable with and set parameters for the technical issues, it could turn its attention to the integration of more disciplines. The papers, under a section titled, “Recording Performance,” reflect this shift, for example, “Reasons for Documenting Gestures and Suggestions for How to go about it” (Seyfeddinipur), “Including music and temporal arts in language documentation” (Barwick), “The language of food” (Pollock), “Botanical collecting” (Conn), “Fieldwork in ethnomathematics” (Chemillier), “Cultural Astronomy for linguists” (Holbrook), “Geography: Documenting terms for Landscape features,” (Turk, Mark, O’Meara, and Stea), Toponymy: Recording and Analyzing Place Names in a language area.” (Nash and Simpson).

The papers included in the 2012 volume definitely speak to how disciplines, other than linguistics, can clearly dovetail with language documentation. However, most

²E-MELD was a five-year project designed to preserve language data and documentation through the development of infrastructure for electronic archives. <http://emel.d.org/>

of these articles describe how fieldwork practices are manifested for other disciplines and might overlap in certain ways within a language documentation project. This provided insights to language documenters but reflects a time, though recent, when the field was still coming to terms with how interdisciplinary projects are conceived and carried out. For example, Barwick (2012:171) justifies the documenting of musical genres appealing to Himmelmann's early call for documenting 'ritual speech events' (1998:179) and also reminds the reader, through Woodbury's words, that "documenters take advantage of any opportunity to record, videotape, or otherwise document instances of language use" (Woodbury, 2003:48). Music seems like an easier stretch for documenting since the methods of data collection should be the same. This contrasts with the efforts to understand what other academic disciplines bring to the field in the way of methods and practices that might intersect or conflict with those of language documenters. McClatchey's paper, "Ethnobiology: Basic methods for documenting biological knowledge represented in language," focuses on techniques for biological data collection and much more. Again, it provides excellent insights into how the documenting of biological information takes shape and that information is certainly helpful to language documenters. But it describes how biologists might do the work along side linguists but not how biologists and language documenters might work together with a single data set or how they might choose to enact the collection of data in a community context. Clearly, the field has embraced Himmelmann's directive to document 'language practices' and has taken to heart the advantages of interdisciplinary work within the context of language documentation. However, there is still room to grow in the direction of achieving integrated interdisciplinary language documentation projects.

With all of this good work coming forward, it is interesting that, at present, even though language documentation encourages an interdisciplinary approach, there has been little discussion about the actual complexities of making that happen. What we cannot glean from the above examples is how interdisciplinary projects are constructed in terms of data collecting, how researchers negotiate more than one field—from theory to method, how decisions are made in the field and afterward in relation to the sharing of outcomes. Further, we need to be more clear about the ethical considerations that come into play when working with other disciplines while engaged with the speaking community(s) involved.

In an effort to get at some of those considerations, a speaker series was sponsored by the National Science Foundation at the 3rd International Conference of Language Documentation and Conservation in Honolulu, Hawaii, in 2013. Four speakers, all language documenters who were engaged in linguistic fieldwork with researchers from four very different disciplines, were asked to reflect on the experience of constructing interdisciplinary projects. The speaker's series featured Jonathan Amith's work on ethnobotany with Nahuatl in Mexico, Birgit Hellwig's study of child language acquisition in Papua New Guinea; Jeff Good's work in Cameroon in partnership with an anthropologist on areal linguistics, and Niclas Burenhult's research on landscape and semantic domains in Malaysia. There were some generalities each researcher discussed:

1. There are sometimes competing goals that cause problems in the field. Burenhult (forthcoming) asks, "How do researchers reconcile the goals of their documentation projects with the both the theoretical and practical goals of the other disciplines involved?"

2. There may be similar goals, but conflicting methodologies. Hellwig (forthcoming) notes that, even though there was a great need for language acquisition studies in the context of language documentation, “Child language studies require experimentation and longitudinal data which are not part of classic language documentation.”
3. It is really necessary to make an effort to understand the partnering researcher and their field to a greater degree than might be expected.
4. Determining how to integrate disciplines in a single project should be established at the outset.

In spite of the challenges, language documentation continues to be a fertile place for interdisciplinary work to grow and researchers in language documentation do understand the advantages. Among them are: 1) The opportunity to bring a different set of research questions to the same project data; 2) Funding agencies realize more research outcomes for their money in such broader-based research; and 3) Multi-faceted project designs which create fuzzy boundaries, which can be a good thing; such designs even create new disciplines.

I believe it is important for language documenters to look more closely at the history, details and application of interdisciplinary work broadly and then bring that understanding into a language documentation framework. A classic definition can help understand how to approach this:

Interdisciplinary research is a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice. (National Academic Press 2004)

Too often, the temptation is to construct a good documentation project and then add a research team or individual colleague from another field to broaden the perspective, perhaps, with the hope of making the project more attractive to funding agencies. Or, sometimes linguists try to go it alone—such as perhaps using a handy botanical guide to fill in the interdisciplinary blanks. Both of these approaches have been tried often, and usually fail to provide a valid picture. Truly strong research projects are conceived of with interdisciplinary perspectives in mind from the beginning. Research questions are mutually conceived. The key concept is always *integration*. Researchers must ask themselves what this truly means as it can be very complex. A well constructed interdisciplinary project integrates both theory and practice from each participating field. Roles for the participating disciplines are well-defined and the outcomes are jointly presented. (If more than one discipline is involved, but the methods and outcomes stand as separate entities, then the project may be ‘multi-disciplinary’ or ‘cross-disciplinary’ but it is not interdisciplinary). Examples of newer approaches are specifically found in articles by Jonathan Amith and Jeff Good in Penfield (forthcoming).

There are three defining aspects of true interdisciplinary research: 1) that the varying disciplines are joined early in the planning process and ‘integrated’ in terms of theoretical and practical input to the targeted research. (There are some exceptions, especially true of language documentation I think, where a second or third discipline might join a project

later when the data reveals the need for it. This was noted by Evans in 2012 when he outlined the ‘strategies for interdisciplinary fieldwork’ (185). 2) that interdisciplinary projects have the potential to form whole new disciplines and, as such, 3) they must, in some way, meet the goals for the research design of each discipline involved. That is, they must adhere to the vision that the central purpose is to solve problems not resolved by one discipline alone. This entails the need to compromise and develop a ‘collaborative personality’ as noted by Jeff Good (forthcoming) as well as any ethical concerns for each discipline. At present, it seems to me, there there is still a need for training of researchers in understanding the parameters and procedures that must govern a true interdisciplinary project. Academic institutions have not done their part in making this possible and, in the future, that needs to change.

3. The Future

“We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline.”
(Popper 1963:88)

Because research questions posed in language documentation frequently “cut right across the borders...,” we can expect that interdisciplinary research in this field will be around for a long time. However, the reality is that researchers are bound to academic institution and while academic institutions often claim to support interdisciplinary work, they are, in fact, structured precisely in ways that make it difficult. Our academic culture is largely based on strong disciplinary boundaries, reinforced by professional societies, institutional hierarchies, and publication sources and requirements.

Rhoten & Parker (2004: 6) write,

“The fact is, universities have tended to approach interdisciplinarity as a trend rather than a real transition and to thus undertake their interdisciplinary efforts in a piecemeal, incoherent, catch-as-catch-can fashion rather than approaching them as comprehensive, root-and-branch reforms. As a result, the ample monies devoted to the cause of interdisciplinarity, and the ample energies of scientists directed toward its goals, have accomplished far less than they could, or should, have.”

This has been the background of interdisciplinary studies, but I do think change is coming. There is a dynamic that must change: there are funding agencies which request and support interdisciplinary projects (understanding their great potential), and there are the researchers who desire to do them with the hope of bridging disciplines in ‘out of the box’ ways and seek support from funding agencies but, in the middle, are institutions which stifle such efforts because of their structure alone. These are often more expensive projects, but they don’t have to be. Small collaborative projects can also be envisioned around very targeted interdisciplinary research questions. In any case, for change to occur, some of the things that will have to be addressed are:

- a) The university ‘silos’ need to change. Russell (1991) writes that before the advent of the modern university in the 1870’s, institutions of higher learning were built on a single discourse model, with a uniform set of values shared between teachers and students. After the modern university, based on the German model, was established,

the academic discourse community became fragmented (21). Academia became "...a collection of discrete communities, an aggregate of competing professional disciplines, each with its own specialized written discourse." (5). It is unrealistic to suggest a change back to a single discourse community but it may be possible to break down some of the silo walls, to create a more fluid communication and working relationship across disciplines.

- b) Journals and other avenues of publication need to also be more willing to publish interdisciplinary research. Most are also constrained by the 'silo' effect. Researchers are challenged by where to publish, how to write (across disciplines) and how to engage audiences from other fields. Since publication is still the basis for success in tenure-track positions, this makes it ill advised to suggest that assistant professors undertake interdisciplinary projects.
- c) A restructuring of the university from the administrative management side This is the only way to begin to bridge disciplines within a given institution. This includes financial support for interdisciplinary programs.
- d) Department-to-department initiatives need to be encouraged. Colloquia, conferences, inter- and intra-departmental events of all types can be used as discussion points for how interdisciplinary research might proceed for everyone's benefit. Much depends on how researchers see themselves in relation to their discipline and how willing they are to push their own limits. Agreements across academic departments or programs can be complicated. Jeff Good comments:

Effective interdisciplinary research often requires collaborators to gain fairly deep knowledge about how practitioners of other disciplines collect and theorize on their data, and may further result in academic outputs that are neither fish nor fowl, as it were, in terms of disciplinary evaluation. Is a culturally informed collection of place names ... an instance of linguistics, anthropology, or geography? Questions like this do not merely provide interesting intellectual puzzles. They can have real-world consequences given the fact that disciplines do not merely exist to provide a convenient way to categorize different methods of inquiry but are also embedded within the institutional structures which support scholarship. (forthcoming)

- e) Training opportunities are needed to teach researchers how to find colleagues with the interest and expertise needed to partner for these projects, how to negotiate the shared responsibilities, how to integrate the relevant areas of theory and practice, how to find funding and where to disseminate outcomes. There may also be theoretical or methodological challenges leading to problems in both the conception of and implementation of the research in question. Language documentation fieldwork carries with it an established methodology for data collection and ethical rules of engagement with community partners which may not be shared or recognized by the participating discipline. The definition of 'fieldwork' itself might differ from that in other disciplines and certainly fieldwork methodologies can differ and become a source of conflict. Issues also tend to arise around data management and ownership. For these reasons specifically, a designated interdisciplinary research team needs to address and anticipate as many

of these things in advance when possible. Ethical considerations also must be addressed across disciplines and in engagement with the speech community. There are training opportunities in place for linguistic fieldwork to begin to address all of these issues, most notably, The Institute on Collaborative Language Research (CoLang), to be held next at the University of Montana in 2020.

The most important consideration is that interdisciplinary projects can take longer to establish, fund and enact. Researchers must be prepared to recognize this time / energy commitment. Finding funding, alone, can be time intensive. It can, however, certainly be worth it. In larger agencies, interdisciplinary projects usually require co-review from the participating programs, adding to the complexity and timing, but possibly also garnering more funding. Smaller, very focused interdisciplinary projects may be fundable through private foundations as well.

In the end, my belief remains that language documentation projects are inherently richer when they take on interdisciplinary characteristics, as reflected in Himmelmann's early vision. Around well-designed research questions, the same data gathered at least doubles in value when it serves more than one discipline. This translates to a richer source of information for researchers but, even more importantly, provides the speech communities with more layers of well-documented aspects of their cultures and languages for posterity.

References

- Amith, Jonathan. forthcoming. Endangered language documentation: The challenges of interdisciplinary research in ethnobiology. In Susan Penfield (ed.), *Interdisciplinary approaches to language documentation in practice*. (Special Issue of Language Documentation & Conservation.) Honolulu: University of Hawai'i Press.
- Barwick, Linda. 2012. Including music and the temporal arts in language documentation. In Nicholas Thieberger (ed.), *The Oxford Handbook of Linguistic Fieldwork*, 166-179. Oxford: Oxford University Press.
- Burenhult, Niclas. forthcoming. Domain-driven documentation: The case of landscape. In Susan Penfield (ed.), *Interdisciplinary approaches to language documentation in practice*. (Special Issue of Language Documentation & Conservation.) Honolulu: University of Hawai'i Press.
- Epps, Patience, Anthony Webster & Anthony Woodbury. 2017. A holistic humanities of speaking: Franz Boas and the continuing centrality of texts. *International Journal of American Linguistics* 83. 41-78.
- Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of Language Documentation*. Berlin and New York: Mouton de Gruyter.
- Good, Jeff. forthcoming. Interdisciplinarity in areal documentation: Experiences from Lower Fungom, Cameroon. In Susan Penfield (ed.), *Interdisciplinary approaches to language documentation in practice*. (Special Issue of Language Documentation & Conservation.) Honolulu: University of Hawai'i Press.
- Evans, Nicholas. 2010. *Dying Words: Endangered languages and what they have to tell us*. Malden, Mass: Wiley-Blackwell.
- Evans, Nicholas. 2012. Anything can happen: The verb lexicon and interdisciplinary fieldwork. In Nicholas Thieberger (ed.), *The Oxford Handbook of Linguistic Fieldwork*, 183-208. Oxford: Oxford University Press.
- Hellwig, Birgit. forthcoming. Child language documentation: A pilot project in Papua New Guinea. In Susan Penfield (ed.), *Interdisciplinary Language Documentation in Practice*. (Special Issue of Language Documentation & Conservation.) Honolulu: University of Hawai'i Press.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161-195.
- Penfield, Susan. forthcoming. Introduction. *Interdisciplinary approaches to language documentation in practice*. (Special Issue of Language Documentation & Conservation.) Honolulu: University of Hawai'i Press.
- Popper, K.R. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Routledge and Kegan Paul.
- McClatchey, Will. 2012. Basic methods for documenting biological knowledge represented in languages. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 281-297. Oxford: Oxford University Press.
- National Academy of Sciences. 2004. *Facilitating interdisciplinary research*. Washington, DC: National Academy Press.
- Rhoten, D. & A. Parker. 2004. Risks and rewards of an interdisciplinary research path. *Science* 306. 2046.
- Russell, David R. 1993. Vygotsky, Dewey, and Externalism: Beyond the Student/Discipline Dichotomy. *Journal of Advanced Composition* 13. 173-94

- Thieberger, Nicholas (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press.
- Woodbury, Anthony C. 2003. Defining Documentary Linguistics. *Language Documentation and Description* 1. 33–51.

Susan Penfield
susan.penfield@gmail.com

Reflections on language community training

Colleen M. Fitzgerald
The University of Texas at Arlington

I reflect upon four decades of language community training, treating Watahomigie & Yamamoto (1992) and England (1992) as the starting point. Because the training activities these papers report began in the 1970s, there is a convincing and growing literature on training, including work published in the years since Himmelmann's (1998) article. The upshot of my reflections is this central point: Language documentation is better when it occurs alongside an active training component. Underlying this point is an acknowledgement that linguists and communities are engaged in mutual training, and in fact, that a binary distinction between linguist and community member is a false dichotomy. The Chickasaw Model, a model that formalizes training, linguistic analysis, documentation, and revitalization as a feedback loop (cf. Fitzgerald & Hinson 2013; 2016), offers a way to capture a fully integrated approach to training. I conclude with nine significant contributions growing out of the training literature.

1. Introduction¹ Linguistics and documentary linguistics benefits from close to half a century of research on training, much of it predating the “official” inauguration of the era of language documentation (Himmelmann 1988). However, the literature on training provides a crucial foundation to many approaches to language documentation and revitalization. In reflecting upon training, have linguists learned anything? What might the consequences and future implications be in training communities and others engaged in documentary projects? I address some of these issues here.

A compelling account of training emerges in two of the papers from Hale et al. (1992), each describing training activities started in the 1970s. One is situated in Arizona in

¹This material is based upon work supported by the National Science Foundation under Grant No. BCS-1263699, “Collaborative Research: Documentation and Analysis of the Chickasaw Verb,” and is also based upon work supported by, and conducted while serving at the National Science Foundation. Any opinions, findings, and conclusions expressed in this material are those of the author, and do not necessarily reflect the views of the National Science Foundation. Thanks to two anonymous reviewers and the editors of this volume for comments on an earlier version of this paper.

the United States (Watahomigie & Yamamoto 1992) and the other in Guatemala (England 1992). Watahomigie & Yamamoto (1992: 12) lay out an early ethics lesson by centering on the responsibility of academics to engage local community members through training:

The goal of collaborative research is not only to engage in a team project but also, and perhaps more importantly, to provide opportunities for local people to become researchers themselves. As Watahomigie & Yamamoto state (1987: 79), 'It is vitally important that anthropologists and anthropological linguists undertake the responsibility of training native researchers and work with them to develop collaborative language and cultural revitalization and/or maintenance programs.'

In the following sections, I outline my assumptions in this paper, and present a short overview of an effective model of training, the Chickasaw Model (Fitzgerald & Hinson 2013). I then outline nine key findings that originate in training activities and that have led to scientific and societal advances. Given the rapid rate of language loss of the world's estimated 7,000 languages, and the scarcity of resources in terms of people, money, and time, a strategic plan for training and community engagement is essential. But it is also important to articulate precisely how and why training is valuable and essential to language documentation.

2. Preliminaries This paper, at the request of the organizers of this volume, references language community training. The responsibility of linguists to communities is addressed in many places (for example, Wilkins 1992; Rice 2006; Fitzgerald 2007b). In a number of these studies, training is recognized as bidirectional, with linguists are trained as much by language communities as linguists train communities. This point is made in numerous places (see for example, Czaykowska-Higgins 2009; Yamada 2007; 2014; K. Rice 2011; Fitzgerald & Hinson 2013; 2016) and is a fundamental premise of my paper. Let me illustrate how this works by drawing from a collaboration in which I am involved, joint work with Joshua Hinson of Chickasaw Nation that focuses on Chickasaw, a severely threatened Muskogean language of southcentral Oklahoma in the United States. We observe that our partnership, between a linguist and the (Indigenous) director of a language revitalization program, involves "educating and training each other, as well as Chickasaw and UTA participants (Fitzgerald & Hinson 2013: 57)." Skills and knowledge transfer are bidirectional and mutual, and goes beyond these the two of us, filtering outward to others in our organizations and in our region.

Acknowledging that training is mutual is especially important because of how these relationships have the potential to enhance the value of the language work for communities, as well as the potential to diminish or even damage that work. Yamada (2007: 262) brings her observations from the Amazon that the results from collaborative language work are also more productive, noting that "[b]y working together, we accomplish much more than either of us could alone." Stenzel (2014: 289), drawing on a participatory language project situated in the Amazon, describes it as having "the potential to contribute to linguistic studies in unexpected ways and to produce data that is *better* in the sense of being richer and more complete," as well as resulting in outcomes better aligned with community goals. Leonard (2017: 20)'s interviews with community practitioners suggest that "[l]anguage work that identifies and legitimises local notions of language, while not a panacea," can improve the range of possible scientific analyses and

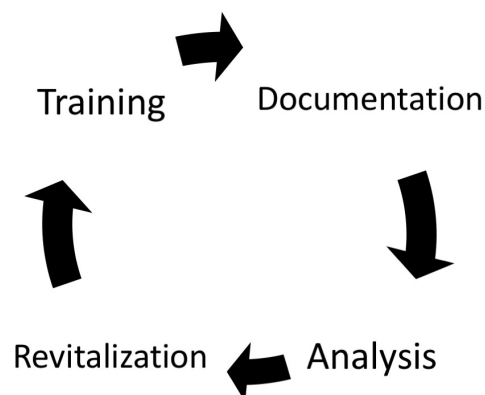
avoid diminishing Indigenous community members' contributions. Better science, better outcomes for community goals, more productive projects are all desirable outcomes. It should be noted that collaborative language work, where training plays a major role, and relationships are essential, should not be viewed without challenges (see, for example, Czaykowska-Higgins 2009; Stenzel 2014). But as noted by all of the researchers cited in this section, these approaches value the expertise held by community members and regard language work as a partnership, one in which linguists are conscious they are also learning.

3. Training alongside documentation, revitalization and linguistic analysis The previous section laid out some assumptions about training and collaborative relationships in language documentation. In this section, I flesh out a more detailed model of precisely how training interacts with language documentation. In this model, a feedback loop is used for the conceptual formulation of the relationship between not just training and language documentation, but also between language revitalization and linguistic analysis.

In a variety of papers, I—along with collaborator Hinson—have argued for framing the collaborative Chickasaw Model (Fitzgerald & Hinson 2013; 2015; 2016; Fitzgerald 2017a; forthcoming) as feedback loop. That is, each stage of language activities produces output which then serves as input to the next stage, creating a mechanism of improvement in response to each stage, feedback occurring as a loop, as in Figure 1. In the Chickasaw Model, language documentation and revitalization are not treated as separate modules, occurring at different times and with no explicit connection. Instead, they are integrated with each other, with revitalization goals driving the documentation and the documentation improving as a result of attention to those goals. Integrating documentation into the revitalization activities improves both kinds of output. When documentation provides rich cultural and linguistic input, it serves as a meaningful learning stimulus, grounded in traditions, conveying community history, and exemplifying the kinds of oral traditions and values often connected with community identity. The Chickasaw Model also predicts how the output of documentation and revitalization will benefit from integrating analysis and training. Analysis of the documentation allows more thoughtful design of revitalization activities, which in turn feeds into training activities that build capacity and engagement within the community. Fitzgerald (2017a) demonstrates how this operates in the context of Chickasaw revitalization, with an eye to phonology and second language acquisition of pronunciation, and Fitzgerald (forthcoming) extends it to activities in the Amazon.

To illustrate this more concretely, I use an example drawn from the community-based language documentation project for Kawaiisu, a highly endangered Uto-Aztecan language of California in the United States. The Kawaiisu Language and Cultural Center has been engaged in a multi-year language documentation and revitalization project, and the team is engaged in transcription and morphemic analysis (Grant & Ahlers forthcoming). These linguistic activities have increased their Kawaiisu language abilities. Elder speakers are producing grammatical items that have been particularly resistant to emerging under elicitation. And these activities in turn are leading to new insights and challenges to earlier analyses of the language by linguists.² Engaging community members in researching their own language, incorporating all four stages of Figure 1, makes advances

²See also Yamada (2007) for another example of how engaging and training native speakers has yielded improved linguistic analyses that challenge prior analyses of the language.



2

Figure 1: The Chickasaw Model (Fitzgerald & Hinson 2013: 59; Fitzgerald 2017a; forthcoming)

in scientific and community goals possible. Both Hale (1965) and Himmelmann (1998) argued for the importance of training and engaging native speakers in research. The Kawaiisu documentary project reinforces the value of training for native speakers and shows its value for second language learners.

4. Nine significant findings from training activities Having laid out the relationship between training and documentation, I now outline nine significant findings that emerge from the training literature.

(1) The documentation (linguistic and otherwise) is richer.

Himmelmann (1998: 176) notes that language documentation should include “as many and as varied communicative events as one can get hold of and manage to transcribe and translate.” Training has expanded the pool of documenters, resulting in turn in richer corpora. For Eastern Chatino of San Juan Quiahije, one such genre occurred in tandem with elections, as town hall oratories were delivered in honor of these events. H. Cruz (2014) analyzes the literary structure of this and other political discourses in her dissertation, accompanied with a documentary collection deposited in the Archive of the Indigenous Languages of Latin America (AILLA) as H. Cruz (n.d.). Her interest in documenting political discourse stems in part from being a diasporic community member, with such events being unfamiliar to her until she returned with a focus on language documentation and orthography development. In the domain of the verbal arts, Fitzgerald (2017c) gives numerous cases where training and revitalization activities enrich resulting documentation. Another testament comes from Linn (2014: 63), who outlines the diversity of language materials that results from a “community-based archive.” Her approach encourages youth involvement whether by new venues online like Facebook or by training Native American youth to use video cameras, and the resulting videos end

up in the language archive. A more diverse set of documenters yields a richer set of language materials in the documentation, such as by providing more context, or adding genres otherwise unnoticed or ignored by academics.

(2) Scientific findings are stronger and more complete.

Growing evidence bolsters the claim that scientific knowledge of a language is enhanced when documentation is collected in environments characterized by mutual training and learning. Certainly, the objects of study may be different when native speakers of Indigenous languages are the linguists and choose research topics. The detail brought to these studies is enriched by the perspective and insights of Indigenous linguists, one reason why Hale (1965) argued that training native speakers as linguists would significantly advance linguistic understanding of language. An example of enriched findings comes from E. Cruz (2017), where she explicates the complexities of naming practices and usage in Quiahije Chatino, an Indigenous language of Mexico. She draws from her own insights as a native speaker and from narratives (cf. the archival deposit, E. Cruz n.d.). In Yamada (2007), which describes her strongly community-centered work with the Kari'ña in Suriname, linguistic training fostered a common meta-vocabulary for talking about language, which in turn strengthens the insights for scientific analyses of the language.

(3) Indigenous community concerns like injustice and trauma affect language work.

Non-Indigenous researchers come from different backgrounds and often do not share the same experiences as Indigenous people, especially regarding violence or trauma, including where language was concerned (cf. Leonard 2017). In many countries, formal education has explicitly or tacitly worked to eliminate minority and Indigenous languages used in homes and communities. In the United States, in my own work, community members have shared their painful stories of boarding or day school where they were punished for speaking their language. Florey (2018) argues it is essential to recognize formal education as a potential barrier, in work done by the Resource Network for Linguistic Diversity in capacity-building and training efforts for Indigenous communities in Australia. Perhaps it is obvious to state, but outsider status can mean there is much one is unaware of, such as political processes and different communicative practices. That lack of knowledge may ultimately be fatal to a project's progress (cf. Fitzgerald 2007a). Privileging the concerns of Indigenous community members increases the potential for more effective and stronger partnerships between communities and non-Indigenous researchers.

(4) Language revitalization and its outcomes are better understood.

The early revitalization literature drew heavily from four language communities: Modern Hebrew, Irish Gaelic, Hawaiian and Māori. From 2000 to 2018, we have seen that literature explode with a host of examples from all over the globe, representing many different contexts. For example, Hawaiian is an Indigenous language of the United States in a location without other Indigenous languages, but other U.S. languages may be found in locales with more than one community language, or no speakers (i.e., sleeping languages), or little documentation. Documentation resources and training can

better support community goals to learn and teach their language as appropriate to that community's linguistic circumstances, such as Breath of Life archival workshops, which have been quite successful in simultaneously serving attendees from distinct sleeping languages at a single venue (Hinton 2001; Fitzgerald & Linn 2013; Sammons & Leonard 2015).

(5) Community reclamation activities are better supported.

Leonard (2012:359) distinguishes language revitalization from reclamation, the former focusing on creation of speakers while the latter is “a larger effort by a community to claim its right to speak a language and to set associated goals in response to community needs and perspectives.” Language reclamation centers community priorities. Returning again to Kari’jna, Yamada (2007) observes that linguistic training of community members empowers their understanding and revitalization of constructions that may have disappeared from wider usage by speakers. Of key importance here is the role that revitalization and reclamation play for a community and its members’ well-being, as well as how those revitalization activities serve as a barometer indicating the vitality of an endangered language (Fitzgerald 2017b).

(6) Language documentation and revitalization training occurs on all continents in order to address local concerns over global language loss.

Delivering training to language community members in the local context is becoming more widespread. Certainly, as Florey & Himmelmann (2009) and others note, training for academics exists (see also Jukes 2011). But training native speakers in descriptive linguistics creates cohorts of Indigenous linguists; this was illustrated by the training Mayan linguists in Guatemala (England 1992). Projects may have a mix of academics (both students and faculty/professional linguists) and language activists together, such as at the Institute on Collaborative Language Research, CoLang (formerly InField), (cf. Genetti & Siemens 2013), as well as in trainings held elsewhere, like those described in Indonesia by Florey & Himmelmann (2009). These approaches are short, perhaps one or two weeks, and the content varies based on the needs of a given locale.

Importantly, short-term training institutes that include language activists are being held for local language communities. Examples exist for local communities of speakers of minority or endangered languages worldwide. In **Asia**, institutes have been held in Tibet (Atshogs et al. 2017, Xun et al. 2017) and Yunnan Province (Mu 2016) in China; in Pakistan by the local Forum for Language Initiatives (Liljegren & Akhonzada 2017); and in Indonesia and Malaysia (Jukes et al. 2017). Training events in **Latin America** have taken place in Mexico (Cruz and Woodbury 2014), in Brazil (Franchetto & Rice 2014; Stenzel 2014; Silva 2016), Peru (Valenzuela 2010; Mihás 2012; Vallejos 2014; 2016), and Guatemala (England 1992; 2003; 2007). In fact, every continent with Indigenous languages has hosted training events: **Europe** (ELAR 2018), **North America** (McCarty et al. 1997; 2001; S. Rice 2011; Fitzgerald & Linn 2013; Fitzgerald 2018a); **Australia** (Amery 2016; Florey 2018, among others); and **Africa** in Ghana (Ameka 2015). And there are examples where training has traversed home locations for both the researchers and the community members, as in the multi-year, multi-location training-based collaboration described for the Kenyan Ekegusii community by Nash (2017). It is worth noting that these examples are drawn from published case studies, but much training occurs without a corresponding publication.

(7) Sustaining language work requires the energy of grassroots community support.

One of the longest running training institutions is the American Indian Language Development Institute (AILDI), which began in 1978 (as described in Watahomigie & Yamamoto 1992). AILDI's ongoing legacy is highly positive, as an empowering site of training, education and language activism for parents, teachers, learners (McCarthy et al. 1997; 2001). It is responsive to community needs, changing the length of the summer session, the offerings, and adding short workshops and offsite training over the last four decades. It was founded and run by key figures in Indigenous language revitalization, tribal citizens Lucille Watahomigie, Ofelia Zepeda and non-Indigenous ally Teresa McCarty. The emergence of two other regional-focused institutes in North America serve similar regional needs: the Northwest Indian Language Institute (NILI, Jansen et al. 2013) and the Canadian Indigenous Languages and Literacy Development Institute (CILLDI, S. Rice 2011). This kind of "Indigenous drive," energy that originates in what communities set as their goals, is analyzed by Fitzgerald (2018a) as essential for sustainable models of language documentation and revitalization.

(8) The most complete understanding of language phenomena draws on Indigenous expertise, ways of knowing and epistemologies.

This point is made insightfully in Leonard (2017), drawing from his analysis of interviews with participants in the Breath of Life and other language workshops. A deeper understanding of the role revitalization plays in a community has drawn on ethnographic and approaches to language revitalization and reclamation (see Granadillo 2006; Meek 2010; Hermes et al. 2012; Davis 2017) that brings different, often Indigenous perspectives to the research questions and investigations. And as argued above (Yamada 2007; 2014; Fitzgerald & Hinson 2013; 2016; Stenzel 2014; Grant & Ahlers forthcoming), stronger and more robust scientific findings are produced by projects that have focused on community-oriented goals in the research project.

(9) Training increases the diversity of linguists and can blur the distinction between linguist and community member.

Expanded opportunities for training, including training and community research involvement are creating pathways into linguistics. Engaging with one's language seems to be that pathway to increasing Native Americans in linguistics (Fitzgerald 2018b), functioning as what some describe as a high impact practice (Kuh 2008), engaging experiences like undergraduate research, internships, and service-learning, all of which positively influence student success, especially for underrepresented students. For example, curiosity over her language, and the non-Indigenous anthropologists "whose job it was to study and describe the lifeways of the O'odham," Tohono O'odham linguist Ofelia Zepeda moved from reading and writing into doing linguistics on her language (Hill & Zepeda 1998: 130).

There are more and more individuals who have roles both as academics and as community members. The range of identities and roles of individuals involved in language work is more complex today than in the 1970s, reported in those foundational papers in Hale et al. (1992). In an annual report on the 2018 California Breath of Life workshop, Hinton (2018: 14) comments on changes in the discipline in linguistics in many domains,

including a heightened awareness of community goals, increasing power over research by Indigenous communities, and "an increase in the number of indigenous people seeking higher education and becoming linguists themselves." More Indigenous scholars are doing linguistics and language work. The body of work they are producing is exciting, asking different questions, integrating different theories, and bringing different perspectives to their languages. In looking at the literature in other disciplines, this should be unsurprising. Outside of linguistics, there is growing evidence supporting a positive correlation between ethnic, racial and gender diversity of teams with an organization's performance (Hunt et al. 2015). And diverse perspectives have been argued to advance science, as Leshner (2011) claims:

increasing the diversity of the scientific human-resource pool will inevitably enhance the diversity of scientific ideas. By definition, innovation requires the ability to think in new and transformative ways. Many of the best new ideas come from new participants in science and engineering enterprises, from those who have been less influenced by traditional scientific paradigms, thinking, and theories than those who have always been a part of the established scientific community.

A critical mass of diversity can itself end up in a feedback loop, fostering the development of more Indigenous linguists as people see role models and colleagues like themselves.

5. Final thoughts It is an uncontroversial point that data from endangered languages has advanced typological and theoretical knowledge in syntax, morphology, phonology semantics and linguistic theories. Importantly, this knowledge production is also occurring in other domains of relevance to linguists. More case studies and analysis of Indigenous language revitalization and training models worldwide will be beneficial, especially if case studies address how these approaches are advancing scientific and other knowledge and testing training models such as the Chickasaw model.

Language documentation is better when it occurs alongside an active training component, and as a result, democratizing training and engaging with communities increases the diversity of linguists. Such important implications are not limited to linguistics; a recent paper in biological anthropology on the ethics of consultation with Indigenous communities over human remains draws on their experiences with training in genomics and supports greater community engagement in order to "produce stronger scientific interpretations and improve relationships between scientists and Indigenous peoples, particularly as the number of Indigenous scientists grows (Bardill et al. 2018: 3)."

Training can effectively be integrated into documentary projects. Additionally, revitalization, in an approach like the Chickasaw Model, serves the community's goals and produces better analysis and documentation of the language. Ultimately, training and engagement with communities results in better science, more diverse scientists, and more empowered communities.

References

- Ameka, Felix K. 2015. Unintended consequences of methodological and practical responses to language endangerment in Africa. In James Essegbey, Brent Henderson & Fiona McLaughlin (eds.), *Language documentation and endangerment in Africa*, 15–35. Philadelphia: John Benjamins.
- Amery, Rob. 2016. *Warraparna Kurna! Reclaiming an Australian language*. Adelaide: University of Adelaide Press.
- Atshogs, Yeshe Vodgsal, Sun Kai, Gnamsras Lhargyal & Chang Min. 2017. The First Tibetan Language and Linguistics Forum. *Journal of Chinese Linguistics*, 45:2. 466–487. (doi:10.1353/jc1.2017.0021)
- Bardill, Jessica, Alyssa C. Bader, Garrison Nanibaa’A, Deborah A. Bolnick, Jennifer A. Raff, Alexa Walker & Ripan S. Malhi. 2018. Advancing the ethics of paleogenomics. *Science* 360: 6387. 384–385.
- Cruz, Emiliana. 2017. Names, naming, and person reference in Quiahije Chatino. In Fernando Armstrong-Fumero & Julio Hoil Gutierrez (eds.), *Legacies of space and intangible heritage: Archaeology, ethnohistory, and the politics of cultural continuity in the Americas*, 163–188. Boulder: University of Colorado Press.
- Cruz, Emiliana. n.d. Chatino landscape collection. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: public. (<https://ailla.utexas.org/islandora/object/ailla%3A124384>) (Accessed 2018-04-25)
- Cruz, Emiliana & Anthony C. Woodbury. 2014. Collaboration in the context of teaching, scholarship, and language revitalization: Experience from the Chatino Language Documentation Project. *Language Documentation & Conservation* 8.262–286. (<http://hdl.handle.net/10125/24607>)
- Cruz, Hilaria. 2014. *Linguistic poetics and rhetoric of Eastern Chatino of San Juan Quiahije*. Doctoral dissertation, University of Texas at Austin.
- Cruz, Hilaria. n.d. Chatino Collection of Hilaria Cruz. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: public. PID ailla:119695, 119853, 119602, 119854 and 119703. (Accessed 2018-04-25)
- Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation* 3. 15–50. (<https://scholarspace.manoa.hawaii.edu/handle/10125/4423>)
- Davis, Jenny L. 2018. *Talking Indian: Identity and language revitalization in the Chickasaw renaissance*. Tucson: University of Arizona Press.
- England, Nora C. 1992. Doing Mayan linguistics in Guatemala. *Language* 68. 29–35.
- England, Nora C. 2003. Mayan language revival and revitalization politics: Linguists and linguistic ideologies. *American Anthropologist* 105. 733–743.
- England, Nora C. 2007. The influence of Mayan-speaking linguists on the state of Mayan linguistics. In Peter K. Austin & Andrew Simpson (eds.), *Endangered languages*, 93–111. Berlin: Helmut Buske Verlag.
- ELAR. 2018. Endangered Language Archive, SOAS, Blog. (<https://blogs.soas.ac.uk/elar/>) (Accessed 2018-05-01)
- Fitzgerald, Colleen M. 2007a. Developing language partnerships with the Tohono O’odham Nation. Working together for endangered languages: Research challenges and social impacts. In Maya Khemlani David, Nicholas Ostler & Caesar Dealwis, (eds.), *FEL Proceedings XI*, 39–46. Bath, England: The Foundation for Endangered Languages.

- Fitzgerald, Colleen M. 2007b. 2006 presidential address: Indigenous languages and Spanish in the United States: How can/do linguists serve communities? *Southwest Journal of Linguistics* 26:1. 1–15.
- Fitzgerald, Colleen M. 2017a. The sounds of Indigenous language revitalization. Paper presented as invited plenary address at the annual meeting of the *Linguistic Society of America*. Austin, TX, January 5–8, 2017. (www.youtube.com/watch?v=wrPe_6Kdo0o&list=PLc4TBef_CiuokIawF2lrXL2q6vyP4IIs7) (Accessed 2017-02-03.)
- Fitzgerald, Colleen M. 2017b. Understanding language vitality and reclamation as resilience: A framework for language endangerment and “loss” (Commentary on Mufwene). *Language*. e280–e297. (https://www.linguisticsociety.org/sites/default/files/e8_93.4Fitzgerald.pdf)
- Fitzgerald, Colleen M. 2017c. Motivating the documentation of the verbal arts: Arguments from theory and practice. *Language Documentation & Conservation* 11. 114–132. (<http://hdl.handle.net/10125/24728>)
- Fitzgerald, Colleen M. 2018a. Creating sustainable models of language documentation and revitalization. In Shannon Bischoff & Carmen Jany (eds.), *Insights from practices in community-based research: From theory to practice around the globe (Trends in Linguistics 319)*, 94–111. Berlin: De Gruyter Mouton. (<https://doi.org/10.1515/9783110527018-005>)
- Fitzgerald, Colleen M. 2018b. Increasing representation of Native Americans in the language sciences and STEM. Paper presented in the *Investing in Diversity Series*, National Science Foundation, Alexandria, VA, March 29, 2018.
- Fitzgerald, Colleen M. Forthcoming. Understanding language documentation and revitalization as a feedback loop. In Stephen Fafulas (ed.), *Amazonian Spanish: Language contact and evolution*. John Benjamins.
- Fitzgerald, Colleen M. & Joshua D. Hinson. 2013. ‘Iittibatoksali ‘We work together’: Perspectives on our Chickasaw tribal-academic collaboration. In Mary Jane Norris, Erik Anonby, Marie-Odile Junker, Nicholas Ostler & Donna Patrick (eds.), *FEL proceedings XVII: Endangered languages beyond boundaries: Community connections, collaborative approaches, and cross-disciplinary research*, 53–60, Bath, England: The Foundation for Endangered Languages.
- Fitzgerald, Colleen M. & Joshua D. Hinson. 2015. Using listening workshops to integrate phonology into language revitalization: Learner training in Chickasaw pronunciation. Paper presented at the *4th International Conference on Language Documentation & Conservation*, University of Hawai‘i, Honolulu, Hawai‘i, 26 February–1 March 2015.
- Fitzgerald, Colleen M. & Joshua D. Hinson. 2016. Approaches to collecting texts: The Chickasaw narrative bootcamp. *Language Documentation & Conservation* 10. 522–547. (<http://hdl.handle.net/10125/24717>)
- Fitzgerald, Colleen M. & Mary S. Linn. 2013. Training communities, training graduate students: The 2012 Oklahoma Breath of Life Workshop. *Language Documentation & Conservation* 7. 252–273. (<http://hdl.handle.net/10125/4596>)
- Florey, Margaret. 2018. Transforming the landscape of language revitalization work in Australia: The Documenting and Revitalising Indigenous Languages training model. In Shannon Bischoff & Carmen Jany (eds.), *Insights from practices in community-based research: From theory to practice around the globe (Trends in Linguistics 319)*, 314–337. Berlin: De Gruyter Mouton.

- Florey, Margaret & Nikolaus P. Himmelmann. 2010. New directions in field linguistics: Training strategies for language documentation in Indonesia. In Margaret Florey (ed.), *The endangered languages of Austronesia*, 121–140. New York: Oxford University Press.
- Franchetto, Bruna & Keren Rice. 2014. Language documentation in the Americas. *Language Documentation & Conservation* 8. 251–261. (<http://hdl.handle.net/10125/24606>)
- Genetti, Carol & Rebecca Siemens. 2013. Training as empowering social action: An ethical response to language endangerment. In Elena Mihas, Bernard Perley, Gabriel Rei-Doval & Kathleen Wheatley (eds.), *Responses to language endangerment. In honor of Mickey Noonan. New directions in language documentation and language revitalization*, 59–77. Amsterdam: John Benjamins.
- Granadillo, Tania. 2006. *An ethnographic account of language documentation among the Kurripako of Venezuela*. PhD dissertation, University of Arizona. (<http://hdl.handle.net/10150/195916>)
- Grant, Laura & Jocelyn C. Ahlers. Forthcoming. Benefits of community-driven documentation focused on interactional language use. *Language Documentation & Conservation*.
- Hale, Kenneth L. 1965. On the use of informants in field-work. *Canadian Journal of Linguistics/Revue Canadienne De Linguistique* 10:2–3. 108–119.(doi: 10.1017/S0008413100005582)
- Hale, Kenneth L., Colette Craig, Nora England, Laverne Masayesva Jeanne, Michael Krauss, Lucille Watahomigie & Akira Yamamoto. 1992. Endangered Languages. *Language* 68. 1–42.
- Hermes, Mary, Megan Bang & Ananda Marin. 2012. Designing Indigenous language revitalization. *Harvard Educational Review* 82:3. 381–402.
- Hill, Jane H. & Ofelia Zepeda. 1998. Collaborative sociolinguistic research among the Tohono O’odham. *Oral Tradition* 13:1. 130–56
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Hinton, Leanne. 2001. The use of linguistic archives in language revitalization. In Leanne Hinton & Kenneth Hale (eds.), *The green book of language revitalization*, 419–423. San Diego: Academic Press.
- Hinton, Leanne. 2018. The 2018 Breath of Life Archival Institute for Indigenous California Languages. Report, August 2018. (<http://files.constantcontact.com/4a41a004201/a03f882f-4c6e-4acb-b602-43321e884148.pdf>) (Accessed 08-29-2018.)
- Hunt, Vivian, Dennis Layton & Sara Prince. 2015. *Diversity matters*. New York: McKinsey & Company. (<https://www.mckinsey.com/~media/mckinsey/business%20functions/organization/our%20insights/why%20diversity%20matters/diversity%20matters.ashx>) (Accessed 2018-05-01)
- Jansen, Joana, Janne Underriner & Roger Jacob. 2013. Revitalizing languages through place-based language curriculum: Identity through learning. In Elena Mihas, Bernard Perley, Gabriel Rei-Doval & Kathleen Wheatley (eds.), *Responses to language endangerment: In honor of Mickey Noonan. New directions in language documentation and language revitalization, Vol. 142*, 221–242. Philadelphia: John Benjamins Publishing.

- Jukes, Anthony. 2011. Researcher training and capacity development in language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 423–446. Cambridge: Cambridge University Press.
- Jukes, Anthony, Asako Shiohara & Yanti. 2017. Collaborative project for documenting minority languages in Indonesia and Malaysia. *Asian and African Languages and Linguistics* 11. 45–56. (<http://hdl.handle.net/10108/89206>)
- Kuh, George. 2008. *High-impact educational practices: What they are, who has access to them, and why they matter*. Washington, DC: Association of American Colleges and Universities.
- Leonard, Wesley Y. 2012. Framing language reclamation programmes for everybody's empowerment. *Gender and Language* 6. 339–367.
- Leonard, Wesley Y. 2017. Producing language reclamation by decolonizing "language". *Decolonizing 'language endangerment' from the ground up, Language Documentation and Description Special Issue on reclaiming languages* 14. 15–36. (<http://www.elpublishing.org/PID/150>)
- Leshner, Alan. 2011. We need to reward those who nurture a diversity of ideas in science. *Chronicle of Higher Education*, March 6. (<http://chronicle.com/article/We-Need-to-Reward-Those-Who/126591/>) (Accessed 2018-05-04)
- Liljegren, Henrik & Fakhruddin Akhuzada. 2017. Linguistic diversity, vitality and maintenance: A case study on the language situation in northern Pakistan. *Multietnica. Meddelande från Centrum för multietnisk forskning, Uppsala universitet* 36–37. 61–79.
- Linn, Mary S. 2014. Living archives: A community-based language archive model. *Language Documentation and Description, Special Issue on Language Documentation and Archiving* 12. 53–67. (<http://www.elpublishing.org/PID/137>) (Accessed 2016-02-06)
- McCarty, Teresa L., Lucille J. Watahomigie, Akira Y. Yamamoto & Ofelia Zepeda. 1997. School-community-university collaborations: The American Indian Language Development Institute. In Jon Reyhner (ed.), *Teaching Indigenous languages*, 85–104. Flagstaff: Northern Arizona University. (<https://eric.ed.gov/?id=ED415058>) (Accessed 2017-5-27)
- McCarty, Teresa L., Lucille J. Watahomigie, Akira Y. Yamamoto & Ofelia Zepeda. 2001. Indigenous educators as change agents: Case studies of two language institutes. In Leanne Hinton & Kenneth L. Hale (eds.), *The green book of language revitalization in practice*, 371–383. Boston: Brill.
- Meek, Barbra A. 2010. *We are our language: An ethnography of language revitalization in a Northern Athabaskan community*. Tucson: University of Arizona Press.
- Mihas, Elena. 2012. Subcontracting native speakers in linguistic fieldwork: A case study of the Ashéninka Perené (Arawak) research community from the Peruvian Amazon. *Language Documentation & Conservation* 6. 1–21.
- Mu, Sophie. 2016. Where in the World is ELAR: ELDP Yunnan Training (<https://blogs.soas.ac.uk/elar/2016/12/15/eldp-yunnan-training/>) (Accessed 2018-05-01)
- Nash, Carlos M. 2017. Documenting Ekegusii: How empowering research fulfills community and academic goals. In Jason Kandybowicz & Harold Torrence (eds.), *Africa's endangered languages: Documentary and theoretical approaches*, 65–186. New York: Oxford University Press.

- Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4. 123–155.
- Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5. 187–207. (<http://hdl.handle.net/10125/4498>)
- Rice, Sally. 2011. Applied field linguistics: Delivering linguistic training to speakers of endangered languages. *Language and Education* 25:4. 319–338.
- Sammons, Olivia N. & Wesley Y. Leonard. 2015. Breathing new life into Algonquian languages: Lessons from the Breath of Life Archival Institute for Indigenous Languages. In *Papers of the Forty-Third Algonquian Conference: Actes du Congrès des Algonquinistes*, 207–224. Albany: SUNY Press.
- Silva, Wilson. 2016. Animating traditional Amazonian storytelling: New methods and lessons from the field. *Language Documentation & Conservation* 10. 480–496. (<http://hdl.handle.net/10125/24703>)
- Stenzel, Kristine. 2014. The pleasures and pitfalls of a ‘participatory’ documentation project: An experience in northwestern Amazonia. *Language Documentation & Conservation* 8. 287–306. (<http://hdl.handle.net/10125/24608>)
- Valenzuela, Pilar M. 2010. Ethnic-racial reclassification and language revitalization among the Shiwilu from Peruvian Amazonia. *International Journal of the Sociology of Language* 202. 117–130. (<https://doi.org/10.1515/ijsl.2010.017>)
- Vallejos, Rosa. 2014. Integrating language documentation, language preservation, and linguistic research: Working with the Kukamas from the Amazon. *Language Documentation & Conservation* 8. 38–65. (<http://hdl.handle.net/10125/4609>)
- Vallejos, Rosa. 2016. Structural outcomes of obsolescence and revitalization: documenting variation among the Kukama-Kukamirias. In Gabriela Pérez Báez, Chris Rogers & Jorge Emilio Rosés Labrada (eds.), *Language Documentation and revitalization in Latin American contexts*, 143–164. Berlin: Walter de Gruyter GmbH & Co KG.
- Watahomigie, Lucille J. & Akira Y. Yamamoto. 1987. Linguistics in action: The Hualapai bilingual/bicultural education program. In Donald D. Stull & Jean J. Schensul, Jean J. (eds.), *Collaborative research and social change: Applied anthropology in action*, 77–98. Boulder, CO: Westview Press.
- Watahomigie, Lucille J. & Akira Y. Yamamoto. 1992. Local reactions to language decline. *Language* 68:1. 10–17.
- Wilkins, David P. 1992. Linguistic research under Aboriginal control: A personal account of fieldwork in Central Australia. *Australian Journal of Linguistics* 12. 171–200.
- Xun, Xiang, Mary Linn, Zoe Tribur, Xuan Guan, Nathaniel Sims, rGyalthar (RJ Erjintou), Yeshe Vosdal Atsok & You-Jing Lin. 2017. Many firsts: First Sino-Tibetan Summer Linguistics Institute. Paper presented at the 4th International Conference on Language Documentation and Conservation, Honolulu, March 2–5, 2017.
- Yamada, Racquel-Maria. 2007. Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation & Conservation* 1. 257–282. (<http://hdl.handle.net/10125/1717>)

Yamada, Racquel-María. 2014. Training in the community-collaborative context: A case study. *Language Documentation & Conservation* 8. 326–344. (<http://hdl.handle.net/10125/24611>)

Colleen M. Fitzgerald
cmfitz@uta.edu

Reflections on funding to support documentary linguistics

Gary Holton

University of Hawai'i at Mānoa

Mandana Seyfeddinipur

Endangered Languages Documentation Programme

Funding for documentary linguistics has changed dramatically over the past two decades, largely due to the emergence of dedicated funding regimes focused on endangered languages. These new regimes have helped to shape and reify the field of documentary linguistics by facilitating and enforcing best practices and integrating archiving into the documentation process. As a result both the pace and quality of documentation have improved dramatically. However, several challenges remain, and additional efforts are needed to ensure the sustainability of funding for language documentation efforts. In particular, more funding needs to be allocated toward training and capacity building in under-resourced regions.

1. Dedicated funding regimes for documentary linguistics The development of documentary linguistics over the past two decades is inextricably tied to the development of dedicated funding regimes which support the collection and organization of language documentation records. Best practices have been codified in grant proposal requirements, and proposal guidelines have likewise been shaped by emerging best practices in language documentation. Perhaps the greatest effect of these dedicated funding regimes has been to increase the valorization of the products of documentation and encourage or even enforce the archiving of those products. While there is great need for linguists to devote more time to language documentation activities, the academic reward system continues to place greater value on publications than on archival collections generated by language documentation activities (cf. Berez-Kroeker et al. 2018). Dedicated funding regimes provide a countermeasure to the established reward systems by incentivizing documentary activities which would not otherwise be highly valued within the academy.

The best-known of these are the large privately-funded schemes such as the Endangered Languages Documentation Programme (ELDP) and the Documentation of

Endangered Languages Program (DOBES) and government-funded regimes such as the United States National Science Foundation Documenting Endangered Languages (NSF-DEL) initiative, but there are many smaller private and public funding regimes as well, such as the Foundation for Endangered Language and the Endangered Languages Fund.¹ A detailed review and typology of the myriad funding regimes is beyond the scope of this chapter. Instead, we focus here on the impact that these programs have had on documentary linguistics, particularly with respect to documentation practices, archiving, and triage. We conclude with a discussion of the sustainability of funding for documentary linguistics.

2. Enforcement of best practices The emergence of dedicated funding regimes has both enforced and facilitated best practices in language documentation. At one time it was common practice for linguists to make use of professional quality recording equipment, such as the Nagra III-NP open reel recorder, released in 1958. This equipment could cost the equivalent of \$10,000 or more in today's dollars and produced high-quality recordings with a relatively long shelf life.² However, by the late 20th century most linguists were making recordings using inexpensive consumer-grade cassette recorders, with little attention to long-term preservation. The digital revolution at the turn of the 21st century provided an impetus for change which was reinforced by the requirements of funders. Dedicated funding regimes proved not only willing to fund higher quality and more expensive recording equipment; they required the use of such equipment, including semi-professional digital video cameras and audio recorders and high quality external microphones. Moreover, not only do these funding regimes provide funding for such equipment, they also provide for or encourage training in the proper use of such equipment. For example, ELDP runs annual trainings for new grantees of ELDP-funded projects. Though it doesn't directly provide such training, the NSF-DEL program regularly funds training workshops such as the Institute on Collaborative Language Documentation (CoLang). Participation in these workshops generally increases the chances of success for an NSF-DEL grant application. Taken together, these efforts have helped to cement an accepted best practice for documentary linguistics.

3. Archiving requirement One of the most tangible effects of the dedicated funding schemes is the enforcement of archiving requirements. While archiving has always been a key part of the language documentation process, it has often been considered to be secondary to the production of descriptive materials such as grammars, dictionaries, and text collections. In the early days of documentary linguistics—prior to the Chomskyan turn—primary materials typically found their way to archives only upon the death of the collector, not upon the completion of the project. New funding regimes now require that archiving is completed as an integral part of the project. In fact, some schemes, including ELDP, view archiving as the primary goal of the funding program. The DOBES and ELDP schemes created digital archives from scratch for the express purpose of housing materials collected by their grantees. Arguably, one of the major contributions of the DOBES and ELDP schemes has been to develop and promulgate an integrated model of language

¹For more information on these funding regimes see the relevant websites: eldp.net, dobes.mpi.nl, www.nsf.gov/funding/pgm_summ.jsp?pims_id=2816, www.ogmios.org, www.endangeredlanguagefund.org.

²For example, the Nagra 4SJ retailed for 11,539 Deutschmarks in 1972, the equivalent of over US\$21,623 in 2018 when adjusted for inflation.

documentation which views archiving as the primary focus and primary outcome of the documentation process. Funding schemes enforce this model by requiring archiving as a primary output and by requiring grantees to develop an archiving plan as part of their project proposals. In the 11 years of its existence the DOBES programme funded 67 projects, most of which resulted in large, multimedia digital archive collections. Since its inception in 2002 ELDP has funded more than 400 projects of varying sizes, most resulting in a digital deposit in the Endangered Language Archive (ELAR). ELDP makes clear the priority placed on archiving. Since 2012 ELDP has enforced progressive depositing by requiring grantees to archive recordings and annotations annually and making disbursement of the next tranche of funding contingent on depositing.

Other funders have followed suit by strongly encouraging archiving as a project outcome. For example, the NSF-DEL program now requires that all applicants include an archiving plan which includes a letter of support from a repository which has agreed to accept the applicant's deposit. In addition, NSF-DEL applications must list the location of archiving in the application summary and discuss "plans and methodology for the sustainable, long-term archiving of all data" in the project proposal (National Science Foundation 2016). Perhaps most significantly from the point of view of funding, all of these funders now recognize archiving as a legitimate expense of language documentation work and allow for the costing of archiving in project budgets. ELDP achieves this by providing dedicated archiving facilities for grantees (through ELAR); NSF-DEL achieves this by allowing archiving to be costed in project budgets. Both approaches recognize that accessioning and sustaining digital language deposits is not without cost. In cases where the funder has not provided a dedicated repository, archiving with an external language archive can consume as much as 8% of project costs (DELAMAN 2014).

We like to think that the motivations for language documentation are altruistic and that linguists create archival deposits because they see value in an integrated model of language documentation and archiving. But given the competing demands on a researcher's time, the archiving incentives and requirements imposed by dedicated funding regimes have provided the impetus for the development of hundreds of new language documentation collections. It should be noted that other types of motivation may also help to incentivize language archiving as well. For example, the University of Hawai'i at Mānoa recently implemented an archiving requirement as part of its graduate program in linguistics (Berez-Kroeker 2015). All PhD candidates are required to submit proof of deposit in writing from the archive director to the dissertation committee before the dissertation can be approved. It is probably too early to tell whether such academic requirements will have the same force as financial incentives imposed by granting agencies.

4. Triage New funding regimes for endangered language documentation have also contributed to the development of a more sophisticated notion of triage for language endangerment. Given limited and finite resources, funding agencies must consider a number of different factors in order to set priorities for which language documentation projects should be funded. Beyond the obvious factor of whether the applicant is actually qualified and has the capacity to complete a proposed project, and whether there is evidence of community participation in the project, funders must weigh factors relating to the urgency of the documentation project. Degree of language endangerment is obviously an important consideration in this decision, but many other factors play a role as well.

Obviously we must prioritize documentation of a language whose last speakers may pass away within the next few years over the documentation of a language with millions of speakers, but often the choice is not so clear. For example, a highly endangered language from a very well-documented family may warrant lower priority for documentation than a relatively viable language isolate with no prior documentation. Documenting the world's linguistic diversity requires an approach which is also informed by our knowledge of genealogical relationships. Similarly, a language with unique and previously undocumented typological features may deserve higher priority for documentation than one with more typical features. Sometimes priorities for language documentation may be based on the contexts which are available for documentation. A relatively vital language may exhibit specialized domains of language use and language knowledge which have been lost in more endangered languages and thus are no longer available for documentation. In such cases the documentation of ethnobotanical knowledge or ritual language, for example, in a less endangered language may be prioritized over the documentation of prosaic language in a more endangered language. Of course, the number of factors involved means there is no single right way to prioritize language documentation efforts, but the need to effectively allocate funding has brought the issue of triage to fore of the language documentation enterprise.

5. Challenges As discussed above, the changes in the funding landscape over the past two decades have on the whole been beneficial for the field of language documentation, facilitating best practices and recognizing archiving as a fundamental part of the documentation process. But the emergence of large, dedicated funding regimes has also brought some challenges. Perhaps the greatest among these is the reinforcement of a distinction between documentation and revitalization/reclamation. Most funders of documentary projects prioritize documentation, often to the exclusion of reclamation efforts. For example, ELDP does not fund revitalization or language maintenance projects, taking a back seat in what can easily be seen as another colonial intervention. The justification for this focus on documentation is understandable, given the urgent need to create a record of languages while they are still spoken. However, this approach has both theoretical and practical limitations. Many language communities view documentation and reclamation efforts as intrinsically related and may even view documentation as a means to language reclamation rather than an end unto itself (cf. Fitzgerald 2017). This is especially true in the North American and Australian contexts. A single-minded focus on documentation is thus antithetical to some indigenous conceptualizations of language. Moreover—and in part as a consequence—in practical terms it is often impossible to separate documentation and reclamation efforts. Is a speaker making a recording in order to create a documentary record of her language or in order that her grandchildren may learn her language? Often there is no clear answer.

Documenters can exploit the fuzziness of this distinction in order to engage in reclamation efforts, but the perception of an artificial distinction between documentation and reclamation remains problematic. On the other hand, the situation is different for sub-Saharan Africa where language reclamation is not inextricably linked to documentation. In other words, community and speakers are happy to engage in documentation but may have no interest in reclamation as the language ecology, the pervasive multilingualism on the ground is the basis for a different language conceptualization (McGill & Austin 2012, Seyfeddinipur 2016, Seyfeddinipur & Chambers 2016).

Documenters can exploit the fuzziness of this distinction in order to engage in reclamation efforts, but the perception of an artificial distinction between documentation and reclamation remains problematic.

Another challenge for new funding schemes is the difficulty of striking a balance between enforcing accepted best practices and encouraging innovation. Given the urgency of endangered language documentation, funders are understandably hesitant to support high-risk projects which employ unproven methodologies. However, enforcement of best practices can go too far, leading to a potential stifling of innovation, as researchers adapt their work to fit the funding requirements rather than innovating new approaches. Funders may be reluctant to support new methods, such as “respeaking” techniques (cf. Reiman 2010), because they fall outside the boundaries of accepted practice. Yet the field desperately needs new, innovative methodologies in order to accelerate progress in language documentation, and developing these new methods may require support for unorthodox and untried approaches.

Another challenge is changing old habits and restrictive conceptualizations of language as a purely auditory phenomenon. Establishing best practices in the use of video in documentation remains a challenge despite the fact that there is thorough theoretical and methodological grounding for language a fundamentally multimodal phenomenon (Floyd 2016, Dingemans 2013, Seyfeddinipur 2012). The value of video to both the scientific record and the community and their descendant is clear. Still, a casual review of deposits in DELAMAN archives reveals a strong reliance on audio, including many audio-only documentations. (This is true even in cases where there are no community restrictions on video use.) It is important to note that this is not a technological problem about how to use a video camera. Video equipment is now low-cost, portable and easy to use, and training in the use of video is readily available. Rather, the problem lies in convincing documenters that these non-auditory features of language are worthy of documentation. This constitutes a major challenge for funders trying to support high quality projects which do create a multipurpose (and multimodal) record.

One of the ironies of documentary linguistics is that linguists in regions which have the greatest threat to linguistic diversity tend to have the least access to funding for endangered language documentation. This exclusion is explicit in many national funding regimes. For example, the US National Science Foundation allocates funding only to scholars based at US institutions. And the DOBES programme required that the project leader have an affiliation with a German host institution. ELDP has been more effective than other regimes by having no restrictions on host institution locations, thus allowing funding world wide. Moreover, ELDP offers Small Grants which have no restrictions on academic qualification, allowing non-academics to apply. Nevertheless, more than 75% percent of ELDP funding has been allocated to institutions in just five countries (see figure 1).

That is not to say that all funding is going to documentation of languages in the US, Europe and Australia; merely that ELDP grants tend to be hosted by institutions in those countries. Some of this funding may actually go to scholars from outside those regions who are studying or pursuing postdoctoral research based at US, European or Australian institutions. Moreover, ELDP has made significant efforts to provide training for scholars outside US, Europe and Australia in order to encourage more competitive applications from those regions and to carry the documentation agenda into the areas where the documented languages are spoken. And national institutions outside the US, Europe and Australia are beginning to prioritize language documentation, as

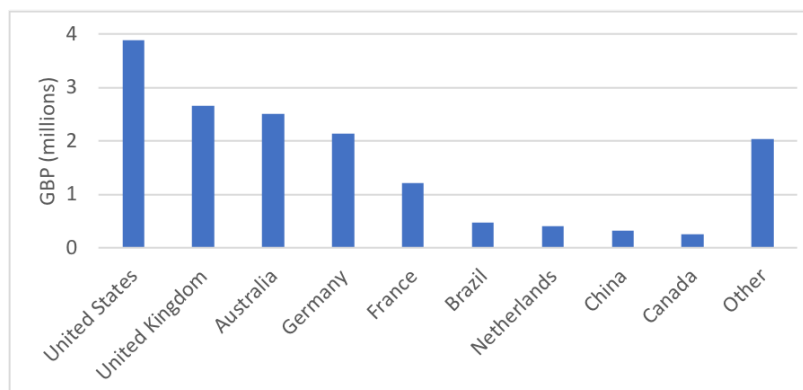


Figure 1: ELDP funding by location of host institution (2003-2016)

evidenced for example by the explicit reference to endangered languages in the 2016 call for proposals by the Indonesian Science Fund.³ Nevertheless, funding for language documentation by scholars without an affiliation in the US, Europe and Australian remains a challenge—a point to which we return below.

6. Sustainability of funding for documentary linguistics Sustainability of funding for language documentation remains an ongoing challenge. While several large funding initiatives have emerged in the past two decades, some of these have already been discontinued. And while these funding programs have accomplished a lot, much more remains to be done. DOBES—once the largest funder of language documentation projects—ended its program in 2011. Throughout the 11 years of its existence the DOBES program was able to fund the creation of archival language documentation corpora for just 67 languages, representing less than one percent of the world’s languages. While most linguists view language documentation as an ongoing effort, private funders typically view the individual projects they fund as finite, with a clearly defined end points.⁴ Government funding bodies face similar constraints, as their constituents would like to see the task of documentation completed. The NSF Documenting Endangered Languages program is regularly threatened by legislative budget cutting. To a certain extent these sustainability issues are a direct result of the way the endangered languages “crisis” has been marketed to potential funders, both private and public. As a crisis, the endangered languages problem should be solvable and time-limited. That is, in characterizing the endangered languages issue as a crisis we have inadvertently defined it as a problem of finite proportions. To a certain extent this is of course true, but in practical terms the documentation of the world’s endangered languages is not likely to be completed quickly.

Even if we could consider a typical 3-year major documentation project to be sufficient for the creation of a documentary collection, current funding regimes are now funding just a handful of such projects per year, perhaps at most 10-20. At that rate it will

³Dana Ilmu Pengetahuan Indonesia. <http://www.dipi.id>

⁴While individual projects may be of defined duration, the funding schemes themselves may be ongoing. Arcadia, the private donor of ELDP just renewed the funding in 2014 with another £7.2 million for five more years as they clearly saw that ELDP was only able to scratch the surface in the twelve years since the programme’s inception.

take at least 150 years to document the slightly less than half of the world's languages which are today considered to be endangered. And during that time even more languages will become endangered, necessitating even more documentation projects. But a typical 3-year documentation project should be considered the bare minimum for creating a record of a language. It is much more common for linguists to devote entire careers to documenting a language. Seen in this light the endangered language documentation "crisis" is not likely to end soon but will rather be an ongoing effort which will need to be carried out over generations. It is unlikely that dedicated funding regimes—whether public or private—will survive over such a long time scale. Hence, linguists will need to look elsewhere to fund language documentation work in the future.

One obvious possibility is to return to the sources which have traditionally funded linguistic research, namely, national funding bodies such as the National Science Foundation (NSF) in the United States or the Deutsche Forschungsgemeinschaft (DFG) in Germany. This approach would have the effect of reintegrating language documentation back into mainstream linguistics. This approach comes with some risk as well. First, the loss of dedicated funding regimes would force language documentation proposals to compete directly with projects in theoretical linguistics, ostensibly leading to less funding for language documentation and to the watering-down of language documentation efforts as applicants introduce theoretical components to their proposals in order to attract funding. Second, the loss of dedicated funding regimes could lead to less enforcement of best practices and less emphasis on archiving. This is especially problematic as many linguistics curricula have not been updated to train the new generations in documentary linguistics theory and methods. As discussed above, dedicated funding has been the primary factor driving the creation of digital multimedia collections as archival deposits; it is not clear that the more general funding schemes would maintain the same strict archiving requirements implemented by EDLP, DOBES, and NSF-DEL. Third, the loss of dedicated funding regimes could lead to less reliance on a pool of expert reviewers who are able to assess the priorities for documentation and the ability of applicants to successfully complete documentation projects. Finally, reliance on more general linguistics funding schemes would almost certainly result in a substantial reduction of funding for endangered language documentation.

A more effective solution to the sustainability problem may be to focus efforts on training local students and scholars, as ELDP has done since 2009. Local linguists will likely have better access to and understanding of the area and the communities. They themselves and their students may be speakers themselves who have a strong interest in documentation and one would hope for a domino effect of training the trainers who in turn train the new generation of documentary linguists working on their own languages. This would also address the issue of not having enough documentary linguists stemming the tide against language loss.


Another approach is to train native speakers to document their own languages. This approach has been advocated by several authors and has motivated the development of regular training institutes for North American languages, such as the American Indian Language Development Institute, the Institute for Collaborative Language Documentation, and the Canadian Indigenous Languages and Literacy Development Institute (Genetti & Siemens 2013, Florey & Himmelmann 2008). Funding training efforts has a number of advantages over funding linguists to do documentation work. The most obvious advantages are economic: native speakers generally do not need to travel in order to engage in documentation activities, so they can commit much greater amounts of time

to the work than can an outside linguist, at a much lower cost. However, even more significant are the quality advantages. Native speakers bring meta-linguistic knowledge which is often inaccessible to outside linguists or else only gained after years of study. Fundamental language documentation tasks such as transcription and annotation are much more effectively completed by native speakers than by outside linguists. Allocating additional funding to training a new generation of language documenters in under-resourced regions will do much more to address the endangered language crisis than will funding allocated to the current generation of documenters in the Global North.

However, training takes time. Developing new capacity in language documentation across the Global South and other under-resourced areas will require a significant commitment, both in terms of time from linguists and in terms of money from funding agencies. It affords hub building to ensure sustained support systems for the communities who are documenting their languages but the return is a major step in the right direction in safeguarding endangered languages. We would do well to turn our attention to language documentation training efforts while funders are still attuned to the endangered language crisis. Once these programs cease, it will be even more difficult to develop this new capacity.

References

- Berez-Kroeker, Andrea. 2015. Reproducible research in descriptive linguistics: Integrating archiving and citation into the postgraduate curriculum at the University of Hawai'i at Mānoa. In Amanda Harris, Linda Barwick & Nick Thieberger (eds.), *Research, Records and Responsibility: Ten years of PARADISEC*, 39–51. Sydney: University of Sydney Press. <http://hdl.handle.net/2123/16674>.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard Meier, Nicholas Thieberger, Keren Rice & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 57(1). 1–18. doi:10.1515/ling-2017-0032.
- DELAMAN. 2014. Report of the DELAMAN costing case study. Accessed 2018-05-11. <http://hdl.handle.net/11122/6928>.
- Dingemanse, M. 2013. Ideophones and gesture in everyday speech. *Gesture* 13. 143–165. doi:10.1075/gest.13.2.02d.
- Fitzgerald, Colleen. 2017. Understanding language vitality and reclamation as resilience: A framework for language endangerment and 'loss' (Commentary on Mufwene). *Language* 93(4). e280–e97.
- Florey, Margaret & Nikolaus P. Himmelmann. 2008. New directions in field linguistics: Training strategies for language documentation in Indonesia. In Margaret Florey (ed.), *The endangered languages of Austronesia*, 121–140. Oxford: Oxford University Press.
- Floyd, Simeon. 2016. Modally hybrid grammar? Celestial pointing for time-of-day reference in Nheengatú. *Language* 92(1). 31–64. doi:10.1353/lan.2016.0013.
- Genetti, Carol & Rebecca Siemens. 2013. Training as empowering social action: An ethical response to language endangerment. In Elena Mihas, Bernard Perley, Gabriel Reid-Doval & Kathleen Wheatley (eds.), *Responses to language endangerment: In honor of Mickey Noonan*, 59–77. Amsterdam: John Benjamins.
- McGill, Stuart & Peter K. Austin. 2012. Introduction and list of contributors (LDD 11). In Stuart McGill & Peter K. Austin (eds.), *Language Documentation and Description*, vol 11., 5–27. <http://www.elpublishing.org/PID/126>.
- National Science Foundation. 2016. Documenting Endangered Languages (DEL). NSF 16-576. Accessed 2018-05-11. <https://www.nsf.gov/pubs/2016/nsf16576/nsf16576.htm>.
- Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4. 254–268.
- Seyfeddinipur, Mandana. 2012. Reasons for documenting gestures and suggestions for how to go about it. In Nicholas Thierberger (ed.), *The Oxford handbook of linguistic fieldwork*, 147–165. Oxford: Oxford University Press.
- Seyfeddinipur, Mandana (ed.). 2016. *African language documentation: new data, methods and approaches*. Honolulu: University of Hawai'i Press. <http://hdl.handle.net/10125/24647>.
- Seyfeddinipur, Mandana & Mary Chambers. 2016. Language documentation in Africa: Turning tables. In Mandana Seyfeddinipur (ed.), *African language documentation: new data, methods and approaches*, 1–7. Honolulu: University of Hawai'i Press. <http://hdl.handle.net/10125/24648>.

Gary Holton
holton@hawaii.edu
 orcid.org/0000-0002-9346-1572

Mandana Seyfeddinipur
ms123@soas.ac.uk

“Because the languages prioritized for documentation are often deeply significant for their speakers as emblems of identity, the movement to study endangered languages has had the salutary effect of rehumanizing linguistics, making it all but impossible to abstract the speakers away regardless of what science might seem to require. [...] Respecting the interests of speakers therefore demands a willingness not only to build relationships across difference, but to approach these relationships analytically.”

(Dobrin & Berson 2011: 207)

Reflections on ethics: Re-humanizing linguistics, building relationships across difference

Ewa Czaykowska-Higgins
University of Victoria

Himmelmann (1998) uses the word ‘ethics’ only once, but his arguments for proposing a field of documentary linguistics reflect assumptions about ethical stances that have been addressed in linguistics publications since 1998. This paper begins by outlining some of these ethical assumptions, and then focuses on considerations closely connected to what Dobrin & Berson (2011: 207) refer to as “re-humanizing linguistics” and “building relationships across difference”. The paper suggests that ethical language documentation work must be grounded in considerations of the human nature of research relationships, the histories of interactions between peoples which inform those research relationships, and varying conceptions of knowledge. Since language documentation work inevitably has social consequences for human beings, aligning language documentation practice with Indigenous research paradigms which emphasize *relational accountability* (Wilson 2008: 99), allows for a practice based on respect, reciprocity and responsibility and ultimately leads to good documentation.

1. Introduction¹ In defining a field of documentary linguistics, Himmelmann (1998) uses the word ‘ethics’ only once, in a list of (sub-)disciplines that, in his words, determine and

¹I am grateful to XwayW’aat Deanna Daniels, Alex D’Arcy, Colleen Fitzgerald, Mary Linn, Keren Rice, and Lorna Wanosts’a7 Williams for discussion, to two reviewers, and to students and colleagues at UVic and at CoLang. I am especially grateful for the generosity of the Elders and knowledge keepers who have worked with me over many years.

influence the “makeup and contents of a language documentation.” In this list, which includes sociological and anthropological approaches to languages, “hardcore” linguistics (theoretical, comparative, descriptive), and language acquisition, he places ethics in a category with language rights and language planning (Himmelman 1998: 167). While Himmelman’s motivations for distinguishing documentary from descriptive linguistics and for providing a format for language documentation do not explicitly name ethical considerations, his arguments implicitly reflect assumptions about ethical stances; many of these have been addressed in linguistics publications since 1998.

In this reflection, I begin by outlining ethical considerations implicitly or explicitly raised in Himmelman. Taking a cue from dictionary definitions of ethics, one could say that thinking about ethics with relation to language documentation involves thinking about how to construct a good language record while acting in a good and right way, informed by the particular social and cultural circumstances of the researcher(s), speaker(s), the language being documented, and its community of practice. My focus here is on issues connected to speakers, communities, and researchers, and what Dobrin & Berson (2011: 207) refer to as “re-humanizing linguistics” and “building relationships across difference”.

The number of researchers doing language documentation work on their own community’s languages is increasing;² however, most published discussions of ethics in language documentation have been written by academically-situated researchers (often from settler-colonial societies like the USA), who are outsiders to the language communities being documented. This paper reflects these limitations: it is informed by my personal and intellectual background, as a Polish-Canadian academic linguist with more than 30 years experience in northwestern North America, including working collaboratively with language communities, activists, knowledge keepers, teachers and learners engaged in language revitalization. Nevertheless, regardless of one’s background, thinking about what it means to build relationships across difference is a necessary part of ethically-informed language documentation work. Ethical work must be grounded in considerations of the human nature of research relationships, the histories of interactions between peoples which inform those research relationships, and conceptions of knowledge. Reflection opens the possibility of acting to shift social relations in research, creating space for multiple perspectives, values and worldviews and thus shaping the documentary linguistic work in ways that might transform institutional practice, decolonize research methodologies (in the sense of Smith 2012) and thus support the production of good documentation.

2. Himmelman (1998) Himmelman begins by writing that “the present reflections have been occasioned in part by the recent surge of interest in endangered languages [...] and the concomitant call for descriptive work on these languages” (1998: 161). The call Himmelman refers to includes work such as that of Hale, Krauss, England, Craig, Watahomigie & Yamamoto (1992). This paper, which appeared in *Language*, was particularly influential in calling attention to the pace at which many languages were undergoing shift as a result of pressure from outside forces (e.g., colonization and globalization) and in making a strong case for the need for a “responsible” linguistics that pays attention to the expectations of linguistics, understood as a science, and to the rights,

²See e.g., Florey 2018 on the DRIL training program in Australia; Fitzgerald & Linn 2013 on Breath of Life workshops in Oklahoma; Zepeda and Hill 1998 on being a community linguist; Hale 1972 on training native speakers in linguistics.

needs, and agency of the speakers of the languages being documented, especially given the pace of language loss. Thus, England (p. 32) discusses “intellectual, scholarly, and political responsibilities to [an endangered] language and the people who speak it”; Craig (p. 23) discuss issues of self-respect and community empowerment, and commitment to linguistic and cultural rights; Watahomigie & Yamamoto (p. 11) acknowledge the agency of speakers of endangered languages in decisions concerning the languages that they speak and propose the possibility of collaborative models of working with language communities.

Himmelmann himself “[...] presume[s] without further discussion that the interests and rights of contributors and the speech community should take precedence over scientific interests” (1998: 172). In this he echoes Wilkins (1992: 189) who states that “...academic linguistic concerns are never so important that they should be allowed to undermine [...] the rights of the host community.”³ This is an ethical stance, from which it follows that requirements for a good language documentation must include access to and accessibility of the language documentation (including fieldnotes and recordings) and ensuring that the documentation is of use for a variety of purposes, including not only those of language-related disciplines, but also those of speech communities involved in maintenance and (re)vitalization of their languages. In distinguishing documentary from descriptive work, Himmelmann argues that many forms of language descriptions are useful primarily to “grammatically oriented and comparative linguists”; he then argues that language documentation involving collections of primary data “[...] have at least the potential of being of use to a larger group of interested parties. These include the speech community itself [...]” (1998: 163). In Himmelmann’s paper, then, language documentation is seen as being a better (where “better” arguably means “more ethical”) response to language loss than language description, and this, in turn, affects the methods and forms proposed for constructing a good documentary record.

In addition to referring to community interests, Himmelmann (1998) also considers the rights of communities and individuals, including rights to privacy, community control, copyright and ownership, “secrecy”, protection of cultural taboos, and prevention of exploitation. These lead to methodological questions about how to edit, curate and archive documentary materials, including ensuring that archived materials do not violate people’s rights to privacy. He also raises questions about community participation in documentary projects, asking how communities can be “actively involved in the design of a concrete documentation project from the very beginning,” even though community ideas about documentation and outside researcher plans may not always agree; finally, he discusses whether and how outside researchers should be involved in language maintenance work “which may be of greater interest to the community than just a documentation” (1998: 188-189). Himmelmann’s arguments and discussions thus include considerations about the responsibilities of language work and workers to speakers and their communities.

3. The human side of fieldwork Recognizing this human, socially-situated side of fieldwork (and of documentation based on fieldwork) is not new. For example, Samarin, who writes from a relatively researcher-focused perspective, says in his 1967 textbook on linguistic fieldwork:

There are obviously questions of ethics when one assumes a role and states a purpose in a community. [...] A more specific question is this: how

³Wilkins (1992: 189) also states that academic concerns should not undermine the personal ethics and sanity of the fieldworker.

can a linguistic investigator take from the people of a community a vast amount of data—much of it given free, all of it given in good will, with the hope perhaps, that it will do them or their children some good—and use it exclusively in scientific publications which in themselves can serve no practical purpose? To some degree the linguist is obliged to his helpers to meet their expectations. Looking upon linguistics as an “objective” science does not make us less dependent on human beings for its pursuit, nor does it make us less obligated to use our findings for the satisfaction of their desires. How all of this is done is, of course, a matter of personal decision. (p. 16–17)

This quote reflects the extent to which language documentation is dependent on the human element, on relationships between the human beings and communities involved, on knowledge and training, expertise, purpose, needs, expectations, agency and power dynamics. As a result, I would argue, *contra* Samarin, that how language researchers respond to the recognition that language documentation work is dependent on human beings cannot simply be “a matter of personal decision”; I would also suggest that it is not sufficient to assume that linguistics is only and always an “objective” science.

In fact, since Himmelmann (1998), ethical considerations connected to the human side of language documentation work have been amongst the most intensely and extensively discussed issues in the literature on documentary linguistics: how to act ethically and how to undertake ethical research in language documentation are being debated and taught in the literature in linguistics, in workshop settings and institutes (like the Institute for Collaborative Language Research (CoLang)), in classrooms, and in language communities. The beginning quote from Dobrin & Berson (2011) thus reflects an increased focus within writings about the intellectual conceptualization and practice of language documentation since 1998.⁴ This increased focus has been influenced by the move in linguistics, particularly since 1998, away from “the generally counterdocumentary trend” that began in “the mainstream of linguistics” in the 1950s (Woodbury 2011: 167). It has also been influenced by work pointing out that language documentation is not historically, politically, socially or culturally neutral, and is not simply an intellectual act (Czaykowska-Higgins 2009: 33–39). Documentation work often occurs in contexts in which the languages being documented are from small communities, many of them Indigenous, that have been marginalized (e.g., by forms of economic or other types of oppression and colonization). Therefore, language documentation work inevitably has social consequences for human beings and this requires particular attention to ethical positions.⁵

There are at least three categories of questions in the literature on ethical issues in language documentation. The first involves issues relating to rights and access: this

⁴This focus is evident when one examines books about linguistic fieldwork published in the last twenty years, all of which have extensive chapters on ethics, (e.g., Bowerman 2008, Crowley & Thieberger 2007, Chelliah & de Reuse 2010, Sakel & Everett 2012, Tsunoda 2005), as do collections of articles on fieldwork and language documentation: e.g., Thieberger (2011) contains a broad overview of ethical issues in linguistic fieldwork including issues related to “ethics codes, individuals, communities, languages, and knowledge systems” (Rice 2011; see also Rice 2006 on rights and obligations in documentary work).

⁵Thus, Dorian writes that “[s]cientists of many stripes like to consider their undertakings apolitical and their professional activities objective and impartial”, but language work is “inevitably a political act” (Dorian 1993: 575), while Hale says that “[t]he scientific investigation of a language cannot be understood in isolation” (Hale 2001: 76). Dobrin & Berson (2011: 188) also suggest “[...]treating linguistic research not as a value-neutral apprehension of intrinsic facts about human symbolic life, but rather as a historically contingent social activity through which linguistics constitutes itself as a discipline (Latour 2005).”

includes consideration of the form, content and value of ethics protocols (e.g., Rice 2012; van Driem 2016), ownership, control, access, protection, copyright, consent, rights and responsibilities to documentation (e.g., Newman 2012; Dorian 2010; Warner et al. 2007), archival research (Innes 2010), legacy resources (O'Meara & Good 2010), and respecting privacy (Macri & Sacramento 2010).

A second, and larger, category of publications is related to the diversity of languages, language contexts and documentary situations, ideologies, people and their responses involved in language documentation work. Thus Wilkins (1992: 188-189) writes "It is important to realise that the social, cultural, political, physical, and historical contexts in which linguists do fieldwork are probably more remarkable for their differences than their similarities. Just as remarkable is the diversity of people who undertake linguistic fieldwork". This point is echoed by Woodbury (2011: 159), who says that "above all, humans experience their own and other people's languages viscerally and have differing stakes, purposes, goals and aspirations for language records and language documentation." Couzens & Eira (2014: 314) point out that linguistics work in communities is "deeply cross-cultural, requiring as it does a productive understanding and connection between sets of ideologies formed for very different purposes, from within very different social and intellectual heritages." Consequently, this diversity and the uniqueness of each documentary situation requires that documentation work must be appropriate to each particular situation (Dobrin et al. 2009: 46-47; Macri 2010).

Perhaps the most debated and discussed theme related to ethics in language documentation in the last 20 years, however, is connected to questioning the extent to which language documentation does or does not require collaborative forms of research, particularly when collaboration involves linguists from outside the language community working with community member language speakers. It is impossible here to do justice to all the discussions focused on this topic (the 2018 *Trends in Linguistics* volume edited by Bischoff & Jany on community-based research provides an excellent overview). These discussions have included sustained arguments for collaborative forms of language documentation work as responses to the Hale et al. (1992) call for a responsible linguistics with social justice as one of its goals. They also have included the defining and elaboration of particular collaborative models such as Participatory Action Research, Community-Based (Language) Research, Empowerment Research, etc. (e.g., Benedicto et al. 2002; Dwyer 2006; Yamada 2007; Penfield et al. 2008; Czaykowska-Higgins 2009; Leonard & Haynes 2010). Studies that illustrate collaboration on particular projects (e.g., Linn et al. 1998; Guérin & Lacrampe 2010; Cruz & Woodbury 2014; Akumbu 2018; Heaton & Xoyon 2018; Junker 2018; other articles in Bischoff & Jany 2018) have also been published, as have discussions of more specific aspects of work involving (outsider) linguists and communities: for example, Stebbins (2012) discusses the differing roles that linguists play within the language communities and universities where they work; Gerdts (2010) discusses the role of the linguist in language revitalization programs; Rice (2011) considers relations with communities in documentary work; Benedicto (2018) lays out challenges associated with aligning the values of collaborative research with those of the academy. Much of this literature sees ethical language documentation work as needing to meet the demands of both documentation and revitalization (e.g., Grenoble 2009: 65) and therefore as requiring community involvement. For Dobrin & Berson (2011: 187) "[...] contemporary documentation in linguistics can usefully be thought of as a kind of social movement" which is "increasingly applied, cognizant of context, and committed to social good". Dobrin & Berson see this social movement as having brought academic

linguists “into a shared space with communities of speakers, researchers working in other disciplines and non-academic institutions, and the public at large”. In this shared space, power imbalances have the potential to be addressed and redressed and the roles of researcher and researched can be re-made. One way to do this is by considering to what extent one’s research practice reflects what Canadian Indigenous scholars refer to as the *Rs*: *respect* between the actors; some measure of *reciprocity* and sense of *responsibility* to the other; the extent to which language documentation is of *relevance* to non-academic as well as academic audiences; and, whether the research involves some element of trust and accountability to *relationships* (e.g., Kirkness & Barnhardt 2001; Wilson 2008; see also Rice 2012, 2018; Czaykowska-Higgins et al. 2018).

Critiques of collaborative models for language documentation have focused on the fact that much of the work on collaborative endeavours in language documentation comes out of societies in which colonization took the form of settlement (particularly, in North and South America, Australia, New Zealand). The most significant point emerging out of critiques is that every language documentation situation is different and therefore collaboration is not necessarily appropriate in every context (e.g., Childs et al. 2014; Crippen & Robinson 2013; Dobrin 2008; Stenzel 2014; Good 2012; Holton 2009). However, if one conceives of collaboration in research, not as a methodology but rather as a philosophy of, or an orientation to research (Ferreira & Gendron 2011: 154 in Rice 2018: 32), and if one assumes that at the heart of all language documentation lies the question of how to enact mutual respect for people, places, relationships, differences in goals, approaches, and methods, then this research orientation allows for multiple ways of understanding collaboration, and thus multiple ways of implementing a philosophy that values human beings and the building of healthy research relationships across difference.

4. What comes next? As mentioned above, Himmelmann’s (1998) motivations for proposing a field of documentary linguistics include arguments for the urgency to document languages before they are “lost forever”. Unquestionably this moral focus on the seriousness of pressures facing languages has brought unprecedented attention in the public sphere, in public national policy decisions, and internationally in such fora as United Nations, UNESCO and the European Union to the need to support languages. This attention has also resulted in increased funding for language documentation. Nevertheless, the rhetoric of endangerment’s focus on language death, loss and weakness (cf., Hill 2002, Errington 2003, Moore 2006, and Perley 2012) runs the risk of perpetuating a colonial narrative in which language speakers and their communities are being acted upon by outside forces and have no agency. An endangerment narrative thus has the potential to reinforce, rather than transform or dismantle, power imbalances between academic outsider researchers and language speakers and their communities. Moreover, the strength and momentum of the language revitalization and reclamation movement (Leonard 2011, 2017) and its relationship, for instance, to the Indigenous human rights movement contrast with a deficit model of language shift and provide arguments in favor of an understanding of language vitality that focuses on resilience (Fitzgerald 2017). In addition, attention to language endangerment, with its focus on enumerating “dying” languages and its understanding of languages as entities which can be objectively delineated and which correspond directly to delineated communities, has been argued to be inappropriate in situations where language boundaries are fluid and where there are high degrees of multilingualism, as in parts of Africa (see, for instance, Ngué Um 2015, Lüpke 2017, Childs et al. 2014, Di Carlo 2016, DiCarlo & Good 2017; cf. Makoni &

Pennycook 2006). In the next few years, therefore, discussion of language endangerment, vitality and ecologies (cf. *Language* 93(4) issue) is likely to influence thinking about ethical engagement in language documentation and about appropriate documentary methodologies around the world.

A second discussion likely to shape the forms of documentations is related to questions about what constitutes linguistic knowledge, how language and its study can/should be understood, and how to bring together Euro-American knowledge systems, epistemologies, and worldviews with those that are not found in the Euro-American academic establishment (e.g., see Rice 2012, Dwyer 2010, Eira & Stebbins 2008, Dobrin & Berson 2011). In relation to Indigenous knowledges, for instance, Leonard (2017: 19) makes the case, following Stebbins (2014) that “[...] Western ideas of language work inherently become elevated over Indigenous ideas when they are uncritically adopted as self-evident, explanatory, and/or accurate.” He further argues that “[...] even well-intentioned Indigenous language work will perpetuate colonial power structures when its products demote ideas from Indigenous communities relative to those of the Western academy, a process Smith (2012: 62) describes as ‘establishing the positional superiority of Western knowledge’” (Leonard 2017: 20). In language documentation contexts there is thus a challenge to colonial mindsets and systems of power in academic research as well as more generally (cf., Couzens & Eira 2014; Kovach 2009; Land 2015; Leonard 2017; Smith 2012; Wilson 2008). Consequently, ethical considerations in language documentation require thinking about, re-examining, and expanding conceptions of language and science, and of what and how to document.⁶

Thinking about how to document, Dobrin & Schwartz (2016) propose that, to conduct socially responsible documentary language research, it might be useful to use the anthropological method of *participant observation*,⁷ which acknowledges the centrality of social relations to research practice from a Euro-American anthropological perspective. Another way to think about social relations in language research, however, would be to align language documentation practice with Indigenous research paradigms which emphasize *relational accountability*, defined in Wilson (2008: 99), who says “[...] methodology needs to be based in a community context (be relational) and has to demonstrate respect, reciprocity and responsibility (be accountable as it is put into action).” From this perspective, the relational is not seen as a bias, but rather, “[...] the relational is viewed as an aspect of methodology” (Kovach 2008: 41). Thus, rehumanizing linguistics, acknowledging the centrality of relationships and difference in language documentation work, emphasizing accountability to those relationships, and grounding ethical research methodologies in social relations is one way that documentary linguistics can continue to move towards de-colonizing, transformative practice.

⁶Wesley Leonard, Megan Lukaniec and Adrienne Tsikewa, three Native American linguists organized a Linguistic Society of America-Society for the Study of Indigenous Languages of the Americas Workshop in 2018 entitled “Expanding Linguistic Science by Broadening Native American Participation.” The workshop, which brought together 40 Native American linguists, community scholars and non-Native linguists, focused on “identifying, valorizing, and disseminating the intellectual tools and cultural values of [language] communities as a way to improve linguistic science.”

⁷Defined as “[...] a research method that is designed specifically to deal with the interpersonal nature of fieldwork in the human sciences [...] that] ties knowledge production directly to the development of social relations.” Dobrin & Schwartz (2016: 253)

References

- Akumbu, Pius W. 2018. Babanki literacy classes and community-based language research. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 266-279. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Benedicto, Elena. 2018. When Participatory Action Research (PAR) and (western) academic institutional policies do not align. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 38-65. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Benedicto, Elena, Dolores Modesta & Melba McLean. 2002. Fieldwork as a Participatory Research activity: The Mayangna linguistic teams. *Berkeley Linguistics Society* 28. 375-386.
- Blommaert, Jan. 2009. Language, asylum, and the national order. *Current Anthropology* 50(4). 415-441.
- Bowern, Claire. 2008. *Linguistic Fieldwork: A practical guide*. Houndmills, England: Palgrave Macmillan.
- Couzens, Vicki & Christina Eira. 2014. Meeting Point: Parameters for the Study of Revival Languages. *Proceedings of The British Academy* 199: 313-334.
- Chelliah, Shobhana & Willem de Reuse. 2010. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.
- Childs, Tucker, Jeff Good & Alice Mitchell. 2014. Beyond the ancestral code: Towards a model for sociolinguistic language documentation. *Language Documentation & Conservation* 8. 168-191.
- Crippen, James & Laura C. Robinson. 2013. In defense of the lone wolf: Collaboration in language documentation. *Language Documentation & Conservation* 7. 123-135.
- Crowley, Terry & Nick Thieberger. 2007. *Field Linguistics: A beginner's guide*. Oxford: Oxford University Press.
- Cruz, Emiliana & Anthony C. Woodbury. 2014. Collaboration in the context of teaching, scholarship, and language revitalization: Experience from the Chatino language documentation project. *Language Documentation & Conservation* 8. 262-286.
- Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation* 3(1). 15-50.
- Czaykowska-Higgins, Ewa, Xway'Waat Deanna Daniels, Tim Kulchyski, Andrew Paul, Brian Thom, S. Marlo Twance & Suzanne C. Urbanczyk. 2018. Consultation, relationship and results in community-based language research. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 66-93. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Di Carlo, Pierpaolo. 2016. Multilingualism, affiliation and spiritual insecurity. From phenomena to processes in language documentation. In Mandana Seyfeddinipur (ed.), *African language documentation: new data, methods and approaches*, 71-104. (Language Documentation & Conservation Special Publication No. 10). Honolulu: University of Hawai'i Press.
- Di Carlo, Pierpaolo & Jeff Good. 2017. The vitality and diversity of multilingual repertoires: Commentary on Mufwene. *Language* 93(4). e254-e262.

- Dobrin, Lise M. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment from Melanesia. *Language* 84(2). 300-325.
- Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. *Language Documentation and Description* 6. 37-52.
- Dobrin, Lise M. & Josh Berson. 2011. Speakers and language documentation. In Austin, Peter K. & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 187-211. Cambridge: Cambridge University Press.
- Dobrin, Lise M. & Saul Schwartz. 2016. Collaboration or participant observation? Rethinking models of 'linguistic social work'. *Language Documentation & Conservation* 10. 253-277. <http://hdl.handle.net/10125/24694>
- Dorian, Nancy C. 1993. A response to Ladefoged's other view of endangered languages. *Language* 69(3). 575-579.
- Dorian, Nancy C. 2010. Documentation and responsibility. *Language and Communication* 30. 179-185.
- Dwyer, Arianne. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 31-66. Berlin/New York: Mouton de Gruyter.
- Dwyer, Arianne. 2010. Models of successful collaboration. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language Documentation: Practice and values*, 193-212. Amsterdam/Philadelphia: John Benjamins Publishing Co.
- Eira, Christina. 2008. Linguists and communities: Discursive practice and the status of collaborative language work in Indigenous communities. *Language and Intercultural Communication* 8(4). 278-297.
- Eira, Christina & Tonya N. Stebbins. 2008. Authenticities and lineages: Revisiting concepts of continuity and change in language. *International Journal of the Sociology of Language*. 1-30.
- Errington, Joseph. 2003. Getting language rights: The rhetorics of language endangerment and loss. *American Anthropologist (Special Issue: Language Politics and Practices)* 105. 723-732.
- Florey, Margaret. 2018. Transforming the landscape of language revitalization work in Australia: The Documenting and Revitalising Indigenous Languages training model. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 314-338. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Ferreira, Maria Pontes & Fidji Gendron. 2011. Community-based participatory research with traditional and indigenous communities of the Americas: Historical context and future directions. *International journal of critical pedagogy* 3(3). 153-268.
- Fitzgerald, Colleen M. 2017. Understanding language vitality and reclamation as resilience: A framework for language endangerment and 'loss' (Commentary on Mufwene). *Language* 93(4). e280-e297.
- Fitzgerald, Colleen M. & Mary S. Linn. 2013. Training communities, training graduate students: The 2012 Oklahoma Breath of Life Workshop. *Language Documentation & Conservation* 7. 252-273.
- Gerds, Donna. 2010. Beyond expertise: The role of the linguist in language revitalization programs. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language documentation: Practice and values*. 173-192. Amsterdam/Philadelphia: John Benjamins Publishing Co.

- Good, Jeff. 2012. 'Community' collaboration in Africa: Experiences from Northwest Cameroon. *Language Documentation and Description* 11. 28–58.
- Grenoble, Lenore A. 2009. Linguistic cages and the limits of linguists. In John Reyhner & Louise Lockard (eds.), *Indigenous Language Revitalization: Encouragement, guidance and lessons learned*, 61–69. Flagstaff, AZ: Northern Arizona University, College of Education.
- Guérin, Valérie & Sébastien Lacrampe. 2010. Trust me, I am a linguist! Building partnership in the field. *Language Documentation & Conservation* 4. 22–33.
- Hale, Ken, Michael Krauss, Lucille Watahomigie, Akira Yamamoto, Colette Craig, Jeanne LaVerne & Nora England. 1992. Endangered languages. *Language* 68. 1–42.
- Hale, Kenneth. 2001. Ulwa (Southern Sumu): the beginnings of a language research project. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 71–101. Cambridge: Cambridge University Press.
- Heaton, Raina & Igor Xoyon. 2018. Collaborative research and assessment in Kaqchikel. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 228–245. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Hill, Jane H. 2002. 'Expert rhetorics' in advocacy for endangered languages: Who is listening, and what do they hear? *Journal of Linguistic Anthropology* 12(2). 119–133.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Holton, Gary. 2009. Relatively ethical: A comparison of linguistic research paradigms in Alaska and Indonesia. *Language Documentation & Conservation* 3(2). 161–175. <http://hdl.handle.net/10125/4424>
- Kirkness, Verna. J. & Ray Barnhardt. 2001. First Nations and Higher Education: The Four R's Respect, Relevance, Reciprocity, Responsibility. In R. Hayge, & J. Pan (eds.), *Knowledge Across Cultures: A Contribution to Dialogue Among Civilizations*. Hong Kong, Comparative Education Research Centre, The University of Hong Kong.
- Innes, Pamela. 2010. Ethical problems in archival research: Beyond accessibility. *Language and Communication* 30. 198–203.
- Junker, Marie-Odile. 2018. Participatory action research for Indigenous linguistics in the digital age. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 164–175. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Kovach, Margaret. 2009. *Indigenous methodologies: Characteristics, conversations, and contexts*. Toronto: University of Toronto Press.
- Land, Claire. 2015. *Decolonizing solidarity: Dilemmas and directions for supporters of Indigenous struggles*. London, UK: Zed Books.
- Leonard, Wesley Y. 2011. Challenging "extinction" through modern Miami language practices. *American Indian Culture and Research Journal* 35(2). 135–160.
- Leonard, Wesley Y. 2017. Producing language reclamation by decolonising 'language'. *Language Documentation and Description* 14. 15–36. .
- Leonard, Wesley Y. & Erin Haynes. 2010. Making "collaboration" collaborative: An examination of perspectives that frame linguistic field research. *Language Documentation & Conservation* 4. 268–293.
- Linn, Mary, M. Berardo & Akira Yamamoto. 1998. Creating language teams in Oklahoma Native American communities. *International Journal of the Sociology of Language* 132. 61–78.

- Lüpke, Friederike. 2017. African(ist) perspectives on vitality: Fluidity, small speaker numbers, and adaptive multilingualism make vibrant ecologies (Response to Mufwene). *Language* 93(4). e275–e279.
- Macri, Martha. 2010. Language documentation: Whose ethics? In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language documentation: Practice and values*, 37–49. Amsterdam/Philadelphia: John Benjamins Publishing Co.
- Macri, Martha & James Sarmiento. 2010. Respecting privacy: Ethical and practical considerations. *Language and Communication* 30. 192–197.
- Makoni, Sinfree & Alastair Pennycook (eds.) 2006. *Disinventing and reconstituting languages*. Clevedon: Multilingual Matters.
- Moore, Robert E. 2006. Disappearing, Inc.: Glimpsing the sublime in the politics of access to endangered languages. *Language and Communication* 26. 296–315.
- Mufwene, Salikoko. 2017. Language vitality: The weak theoretical underpinnings of what can be an exciting research area. *Language* 93 (4). e202–e223.
- Musgrave, Simon & Nick Thieberger. 2006. Ethical challenges in documentary linguistics. Keith Allan (ed.) *Selected papers from the 2005 conference of the Australian linguistics society*.
- Newman, Paul. 2012. Copyright and other legal concerns. In Nicholas Thieberger (ed.), *Oxford Handbook of Linguistic Fieldwork*, 430–456. Oxford: Oxford University Press.
- Ngué Um, Emmanuel. 2015. Some challenges of language documentation in African multilingual settings. In James Essegbey, Brent Henderson & Fiona McLaughlin (eds.), *Language Documentation and Endangerment in Africa*, 195–212. Amsterdam: John Benjamins.
- O’Meara, Carolyn & Jeff Good. 2010. Ethical issues in legacy language resources. *Language and Communication* 30. 162–170.
- Penfield, S., A. Serratos, B.V. Tucker, A. Flores, G. Harper, J. Hill, Jr., & N. Vasquez. 2008. Community collaborations: Best practices for North American Indigenous language documentation. *International Journal of the Sociology of Language* 191. 187–202.
- Perley, Bernard C. 2012. Zombie linguistics: experts, endangered languages and the curse of undead voices. *Anthropological Forum: A journal of social anthropology and comparative sociology* 22(2). 133–149.
- Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4. 123–155.
- Rice, Keren. 2010. The linguist’s responsibilities to the community of speakers: Community-based research. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language Documentation: Practice and values*, 25–36. Amsterdam: John Benjamins.
- Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5. 187–207.
- Rice, Keren. 2012. Ethical Issues in Linguistic Fieldwork. In Nicholas Thieberger (ed.), *Oxford Handbook of Linguistic Fieldwork*, 407–429. Oxford: Oxford University Press.
- Rice, Keren. 2018. Collaborative research: Visions and realities. In Shannon Bischoff & Carmen Jany (eds.), *Insights from Practices in Community-based Research: From theory to practice around the globe*, 13–37. (Trends in Linguistics. Studies and Monographs 319.) Berlin/Boston: De Gruyter Mouton.
- Robinson, Laura C. & James Crippen. 2015. Collaboration: A reply to Bower & Warner’s reply. *Language Documentation & Conservation* 9. 86–66.
- Sakel, Jeanette & Dan Everett. 2012. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.

- Samarin, William. 1967. *Field Linguistics*. New York: Holt, Rinehart & Winston.
- Smith, Linda T. 2012. *Decolonizing Methodologies: Research and Indigenous peoples* (2nd ed.). London, UK: Zed Books.
- Stebbins, Tonya. 2012. On being a linguist and doing linguistics: Negotiating ideology through performativity. *Language Documentation & Conservation* 6. 292–317.
- Stenzel, Kristine. 2014. The pleasures and pitfalls of a ‘participatory’ documentation project: An experience in Northwestern Amazonia. *Language Documentation & Conservation* 8. 287–306.
- Tsunoda, Tasaku. 2005. *Language endangerment and language revitalization*. (Trends in Linguistics Studies and Monographs 148.) Berlin/New York: Mouton de Gruyter.
- van Driem, George. 2016. Endangered language research and the moral depravity of ethics protocols. *Language Documentation & Conservation* 10. 243–252.
- Warner, Natasha, Quirina Luna & Lynnika Butler. 2007. Ethics and revitalization of dormant languages: The Mutsun language. *Language Documentation & Conservation* 1(1). 58–76. <http://hdl.handle.net/10125/1727/>
- Wilkins, David. 1992. Linguistic research under aboriginal control: A personal account of fieldwork in central Australia. *Australian Journal of Linguistics* 12(1). 171–200.
- Wilson, Shawn. 2008. *Research is ceremony: Indigenous research methods*. Halifax/Winnipeg: Fernwood Publishing.
- Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.
- Yamada, Racquel-Maria. 2007. Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation & Conservation* 1(2). 257–282.
- Zepeda, Ofelia & Jane Hill. 1998. Collaborative sociolinguistic research among the Tohono O’odham. *Oral Tradition* 13(1). 130–156.

Ewa Czaykowska-Higgins
ecz@uvic.ca

Reflections on diversity linguistics: Language inventories and atlases

Sebastian Drude
University of Iceland
Goethe-Universität Frankfurt

This contribution gives a short overview of “language inventorying”: research aiming at creating comprehensive catalogues and atlases of all the languages in the world, which has seen a boost with the renewed interest in linguistic diversity triggered by the awareness of language endangerment in the 1990s. By focusing on the development of the ISO standard 639 and SIL’s Ethnologue, the main advances and issues in this area are discussed. The overview concludes by presenting the major alternative resources, in particular Glottolog.

The label “diversity linguistics” has been introduced by Martin Haspelmath and others at the Max-Planck-Institute for Evolutionary Anthropology in Leipzig.¹ To my knowledge, it was first used in the context of the final conference of that institute’s department of linguistics (MPI-EVA 2015). Now there exist a number of activities under this label, including a book series “Studies in Diversity Linguistics” (Haspelmath 2014ff). In a broad sense, the term designates those branches of linguistics that show interest in the diversity of languages, their structure and relationship: descriptive linguistics (especially of previously understudied languages, often in a fieldwork setting), language typology, and comparative linguistics. Language Documentation is included in or at least a close neighbor to this group.

In a narrower sense, the term could refer to those studies that are interested in the diversity in itself, aiming in a first step at creating comprehensive catalogues of the world’s linguistic diversity, where all languages are recorded with the information necessary to identify them, and with additional information. We label these studies here “language inventorying”, and this is the focus of this contribution.²

¹I would like to thank my colleagues in the ISO working group TC37/SC2/WG1 on language coding, and some colleagues who provided me with information and assessments in personal comments, in particular Harald Hammarström, Chris Moseley and Martin Haspelmath.

²Sometimes “Language taxonomy” is used, but this would imply genetic classification, which is often, but not necessarily, part of language inventorying.

Among the most basic identificatory pieces of information needed for each language in a language inventory are its name(s) and its geographical distribution, which is why language maps/atlas together with catalogues are the major tools and products of language inventorying. The additional information to be provided for each language is in principle an open-ended set of properties. Common are the genealogical classification and relationships with other languages, the number of speakers and other aspects of the ‘language context’, including their linguistic vitality, how well studied and described they are, and how much documentation (in the modern sense of annotated multimedia collections of language use, cf. Himmelmann 1998; Gippert, Himmelmann & Mosel 2006) is available. It can also include further structural / typological properties.

The tradition of language inventorying goes back to the Renaissance with Gesner’s (1555) “Mithridates” (referring to the famously polyglot antique Persian/Greek king Mithridates VI). With the increasing outreach of the European empires, this work has been continued in the late 18th and early 19th centuries, for instance with the “Catálogo de Lenguas...” by the Jesuit Hervás y Panduro (1800 ff). The perhaps best known work of this kind was published by Friedrich Adelung and his successor Johann Severin Vater (1806 ff); in reverence to Gesner, it was also called “Mithridates”, and contained a detailed presentation of almost 500 languages with text samples (mostly the Christian “Our Father” prayer). These works informed the formation of what later would become the discipline of linguistics, in personalities such as Wilhelm von Humboldt, one of the most distinguished early scholars interested in diversity linguistics.

The 20th century has seen a few scholars interested in language inventorying, beginning with “les langues du monde” by A. Meillet and M. Cohen (1952 and earlier editions back to the 1920s), and perhaps most notably Voegelin & Voegelin (1977). Also, several original regional overview works were compiled, such as Wurm and Hattori (1981) on the Pacific, or Sebeok (1977) on the Americas. Other compilations covered only a selection of major languages, such as Ruhlen (1987). A milestone was the comprehensive atlas of languages edited by Moseley and Asher (1994).

The interest in having a clear picture of the world’s linguistic diversity, and thus in language inventorying, arose again in the 1990s, when the linguistic community at large became aware that this very diversity is vanishing at a worrying pace (best known are the lead articles in *Language* 1992 by Hale and others; most notably Krauss 1992). The same trigger also led to the establishment of language documentation and to the revival of interest in diversity linguistics, after decades of focusing on linguistic theories, the most prominent of which usually abstracted away from inner and outer linguistic diversity, deeming it enough to study English and perhaps a few other languages.

In the 1990s, the most comprehensive and best known catalogue of the world’s languages was the “Ethnologue”, published by the Summer Institute of Linguistics (now ‘SIL International’). The Ethnologue started out in the 1950s (first edition 1951), compiled first by R.S. Pittman, as a checklist, so to speak, for the Wycliffe Bible Translators (WBT), the sister organization and main sponsor of SIL International that aims at translating the bible in all languages of the world. Maps were included in the fourth edition (1953), and the register grew from a few dozen languages and language groups to several thousand, mostly relying on information provided by SIL/ WBT missionary-linguists. B. Grimes took over as general editor around 1970 and transformed it into a general reference work (now published by SIL), systematically filling gaps, including major languages and those where no first-hand knowledge by SIL/ WBT members was available, also by simply including other work such as Voegelin & Voegelin (1977) or Wurm & Hattori (1981). One major

enhancement was the introduction of unique three-letter codes in the 10th edition (1984), allowing to unequivocally refer to languages despite the notorious confusion involving language names. Its 1996 edition, the oldest one which still can be obtained online (Grimes 1996), *Ethnologue* listed 6703 living languages,³ a number which increased comparatively less since then. When Grimes handed the editorship over in the early 2000s, it was around 6800; the latest five (now annual) editions list just around 7100 languages.

The need for referring unambiguously to languages arose together with the new interest in diversity linguistics: now also linguists were in need of ‘checklists’ for statistics on language diversity, on language endangerment, on coverage of description—and of documentation. This need increased with the rapid technological developments: for instance, software companies need to offer localization of their products in ever more languages; international bodies such as the WWW consortium and UNICODE need to refer to individual languages. Also, with the rising language documentation efforts, language archives and similar institutions need to identify the materials they host, for instance for search engines and portals such as the Open Language Archive Community (2003ff) to be able to aggregate information.

An international standard for this purpose is provided by the International Standardization Organization ISO under the number 639. In fact, this is a group of standards. Its first part, now named ISO 639-1, was established in 1967, it contains (now ca. 200) two-letter codes like “en” for English. The second part ISO 639-2 was approved in 1998; it contains three-letter codes for now around 400 major individual languages⁴ and some 70 codes for groups of languages (mostly either genealogical, e.g. “afa” for Afro-Asiatic languages, or geographical, e.g. “cau” for Caucasian languages). This part 2 came from two sources (terminologists and librarians) which were harmonized.⁵

With the new needs to unambiguously refer to all the other languages in the world, ISO approached SIL international in the early 2000s to include the *Ethnologue*’s three-letter codes as ISO 639 part 3, and to serve as the “registration authority” that maintains and updates this part of the ISO standard. In its 15th edition (2005), organized by R. G. Gordon, Jr., the *Ethnologue* had adopted the ISO 639-2 codes for all languages (replacing hundreds of conflicting codes), and in 2007, ISO 639-3 was established based the *Ethnologue* codes. Since then, *Ethnologue* officially follows ISO in the question what counts as a language.

With this move, the ISO 639 standards came under the attention of diversity linguists at large, and received much critique of different kinds (most prominently perhaps by Morey, Post & Friedman 2013; see also the reply Haspelmath 2013). It is worthwhile to discuss some of these issues in detail as several of them are also relevant for other language inventorying efforts.

³There are contradictory numbers. The online version speaks of 6703 languages; Wikipedia (WikiProject Languages 2015) counts 6883 primary language names, and the list of language codes and names distributed together with the 13th edition in 1998 has 7825 lines/entries (Grimes 1998).

⁴‘Major language’ may be interpreted as “ausbau language” (language by development) in the sense of Kloss (1967): varieties that developed a literary standard and serve official functions, among others. (These are among the criteria to receive a code in ISO 639-1 and 639-2.) In this conception, most other languages are “abstand languages” (languages by distance)—a linguistic variety or group of varieties characterized of being not mutually intelligible with any other variety or language. We have no space here to argue that the concept of “language” designates real entities which are more than an ideological construct, even though the identification of individual languages may be tricky, for instance in the case of dialect chains, see below.

⁵In some 20 cases, two synonym codes were admitted. This happened where the terminologists were using codes mostly based on the autonym of the languages (e.g. “eus” for Basque, or “fra” for French) while the librarian’s MARC codes were based on the English name (“baq” for Basque, “fre” for French).

Certain issues concern the institutional setting and use. For instance, government or funding and other agencies can mistake the fact of a language to be listed or not as being an “authoritative” statement of its status or even existence. This is a problem for any such standard and any language inventory, whoever compiled it, once it is widely accepted. Its close connection with the *Ethnologue* and thus, a missionary organization with potentially several agendas, some open, some less so, is seen as critical by many academic linguists, especially in regions where SIL has contributed little to the local academic development but more to the weakening of indigenous cultures, and, consequently, languages through supporting fundamentalist Christian proselytizing (see the contributions by Dobrin et al. 2009). Yet, alternatives to the current setting would require a solid institutional framework, as the revision process is expensive and high technical reliability is needed for ISO 639 to be acceptable.

Other critiques concern the mnemonic character of the three-letter codes which often look like acronyms. This becomes problematic when the language name that can be identified as the base for the code is deemed inappropriate. This is true, but no good solution seems feasible at this point. More than two thirds of the 17,576 possible combinations are taken; they cannot be recycled for reasons of consistency. If all criticized labels were to be exchanged, there is a real risk of running out of codes (new codes hardly get any mnemonic match anyways). ISO certainly will (and arguably can) not get into the merits of appropriate labels, as many political issues are involved—for instance, who is authorized to complain, and who to decide? The downside is that codes may be rejected by language communities. The perhaps most prominent case is the Mapudungun language, many of whose speakers reject the code “arn” which reminds of the earlier name Araucanian, considered offensive. This, however, makes creating a Wikipedia in Mapudungun problematic, as the use of the ISO code is required.

More fundamental critiques concern the delimitation of languages versus dialects or other varieties, or language families, and the heterogeneous criteria (linguistic and socio-political) applied, even more so as this is a dynamic field where languages die and new varieties and eventually languages emerge. The former points question the very feasibility of language inventorying as such. The latter point (dynamics) is not a serious objection if everybody is aware that any such list is bound to regular revision.

It is true that not only languages as defined as comprising all varieties that are (possibly serially) mutually intelligible (abstand languages in Kloss’ sense) are listed in ISO 639-3 (and the *Ethnologue*), and even not only abstand languages and ausbau languages (see footnote 4), but also many dialects. For instance, there are 18 High German (without Yiddish) and 11 Low German (without Dutch and other Low Franconian varieties) entries although most linguists would agree that there are at most two (ausbau) languages (High and Low German) in one single German dialect continuum.⁶ Also in the case of small language communities, different mutually intelligible dialects have repeatedly received separate entries and codes if they are spoken by ethnically / politically separate groups, as they function as an emblem of ethnic identity.

Still, to give up on creating a comprehensive language inventory on such grounds would mean to throw the baby out with the bath water. In fact, even in (at the first glance) complex and confusing situations involving, in particular, dialect continua, the specialists who work with those varieties usually can come to an agreement if they agree on the

⁶One can argue that some varieties spoken by former German emigrants/colonists (such as pdc, Pennsylvania German, or hrx, Hunsrik) or other geographically separated communities (e.g. cim, Cimbrian) have developed into separate abstand languages.

criteria (linguistic, mainly mutual intelligibility, and socio-political). There will always be borderline cases, but the general situation can in principle be improved and eventually solved by (a) making both the linguistic and socio-political criteria for each entry explicit and (b) improving the criteria and conceptual framework, recognizing that the concrete language topology may be more complex than can be captured by the simple conceptual triad “language family – language – dialect”. The only recent consistent attempt known to me to advance at the second front is offered by T. Kaufman with his classifications of South American languages (Kaufman 1990; 2007), going back to earlier proposals by Hockett (1958). In addition to the most common and paradigmatic case of *families / languages / dialects* (some languages being *dialect chains*), he distinguishes *language areas* and their constituent *emergent languages* (between them there are clear boundaries but high intelligibility), and *language complexes* (these are dialect chains with *virtual languages*: subsets functioning as languages). There may be other elaborate conceptions that I am not aware of, but none of these seems to have influenced current works of language cataloging, where the trichotomy of dialect / language / language family are the only concepts used, usually not even sufficiently distinguishing between languages on linguistic vs. languages on socio-political grounds.

In my view, the most interesting questions concern ISO’s (and the Ethnologue’s) factual adequacy. A very detailed assessment with hundreds of individual corrections has been offered by Hammarström (2015). With regard to scientific standards, the most pertinent critique against the Ethnologue is that most of its data is not verifiable, as sources often remain unknown. The second complex concerns the genealogical classification. (This is not a problem of ISO 639-3, which does not engage in classification at all.) Ethnologue has created genealogical trees which are frequently at odds with the best researched historical-comparative work. In Hammarström’s (2015:734) evaluation, the expert-like-ness is only around 30%. In this context, the “authoritative” character of the Ethnologue is indeed potentially dangerous as readers often will not take the trouble to look for expert classifications, and even scientific publications such as the International Encyclopedia of Linguistics (Frawley 2003) just follow the Ethnologue without verifying the merit of individual genealogical trees.

As to the question of identifying and counting languages, central to this essay, the Ethnologue has often been criticized to overcount (mainly by recognizing too many dialects as languages; see the example of German varieties above). Dixon (2012:463–4) even estimates that the number of languages if consistently grouping all mutually intelligible varieties together in one language, one would arrive at a count of ‘a good deal less’ than 4000 languages. Hammarström (2015:733) again provides the arguably best informed estimate of 6497 known languages (solely on the grounds of mutual intelligibility, but dividing larger dialect continua in a number of different languages, see Hammarström (2005)), or between 5,593 and 7,400 languages with a confidence interval of 95%. Of these, around 6000 are living. Conclusively, by purely linguistic criteria, the Ethnologue overcounts indeed, but only by some 10–15%.

Still, there are other problems with the Ethnologue and ISO 639-3. The very relationship between the two is problematic, because when one presents the Ethnologue with shortcomings and mistakes (many are individually listed and substantiated by Hammarström), Ethnologue can point to ISO 639-3 as the authority on which languages to list; Ethnologue is supposedly only following. ISO 639-3, in turn, will point to its change request submission mechanism, which puts the burden of doing the paperwork for correcting each of the mistakes on the specialists or language inventoryists like

Hammarström, and the outcome is uncertain (and the referees or experts that make the decisions for SIL as registration authority are not known)⁷—unless you work closely with the registrar of ISO 639-3 (or the editors of *Ethnologue*?) to get many changes accepted without having to publish the evidence. In addition, the *Ethnologue* now is behind a paywall, which makes much of its information, in particular the language maps, much less accessible (see the discussion by Skirgård 2016). (ISO 639-3 remains openly accessible.)

In view of these limitations, what are the alternatives, in particular from the academic community? Luckily, there are several, albeit each with its scope and problems. Limitations of space prevent me to discuss each of the existing relevant resources that have been created over the last twenty years or so, largely as a result of the renewed focus on diversity linguistics and fieldwork and language documentation efforts.

It is clear that such resources should not be created by one person or small group alone, but needs contributions from many experts. This is well demonstrated by the *linguasphere registry* (Dalby, Barrett & Mann 1999), which shows a resourceful conception, many valid insights and much knowledge, but is preliminary for many individual languages / families / areas (and does not cite any sources) (Vajda 2001). Its 8-level language codes have not been widely taken up, and it has been rejected as the basis for a planned part 6 of ISO 639 on language varieties.

There are strong arguments that suggest that the most advanced and sound catalog of the world's languages is the *Glottolog* (Hammarström et al. 2018; see also Hammarström 2016). It lists all languages with their genealogical and geographical position according to the best information available to the editors.⁸ More importantly, it lists many important sources for each language. On the other hand, at this point *Glottolog* gives no language context information such as speaker numbers or vitality, but they provide a reliable assessment of the degree of description (they plan on including more, see Hammarstrom et al. in prep.), also presented in the comprehensive “Handbook of Descriptive Language Knowledge” (Hammarström 2018).

Glottolog assigns codes (the ‘Glottocode’) of letters and ciphers, which are being taken up by important other resources such as Wikipedia. The authors were partly also involved in creating other resources important for diversity linguists, in particular *WALS* (Dryer & Haspelmath 2013). *Glottolog* is currently actively maintained, but does not seem to have a sustainable long-term institutional basis; as happens with most such resources: if the current editors decide to stop maintaining it, there is no guarantee that it will be kept alive. Nevertheless, at this point it is arguably the most complete and reliable language catalogue, surpassing other works in particular in its genealogical classification and balanced judgment as to identify languages versus dialects, but it needs to be complemented with other sources for language context information.

Two further major resources are also well-researched and do provide language context information, but are restricted to endangered languages. The *UNESCO Atlas of the World's Languages in Danger* (Moseley 2010) was first edited as a book (Wurm 1996), has been largely enhanced in the early 2000s and is still updated by its main editor, although the funding currently barely allows to keep the atlas accessible. Chris Moseley is also co-editor of the *Routledge Atlas of the World's languages* (Moseley & Asher 2007), itself

⁷Besides this problem, the transparency has increased a lot since 2007; for a few years now all decisions are publicly visible. In the 2018 round of change requests, anyone can see and comment the requests, which improves the possibility for experts and community members to interfere considerably.

⁸Based on the geographical information (unfortunately, just a coordinate, the geographical ‘centre-point’) it is possible to create a world map of the languages according to *Glottolog*, see Caines et.al. (2016; realized as Caines & Bentz).

arguably one of the most complete language catalogues, although its reliability varies among different areas.

UNESCO is planning on a new, much more comprehensive and ambitious edition including all languages. The UNESCO atlas applies its own endangerment measurement based on its experts panel's conception (UNESCO Ad hoc Expert Group on Endangered Languages 2003), which is different from (and not rarely comes to different results than) Ethnologue's EGIDS scale (Lewis & Simons 2010).

The Endangered Languages Project (ELP) (Alliance for Linguistic Diversity 2013ff) has set up a website where users can contribute with information and language resources such as recordings. The underlying technology has been provided by Google.org, the research by NSF grants led by the LinguistList and the University of Hawaii. Underlying the ELP is ELCat, the Endangered Languages Catalog, developed mainly at the University of Hawaii. It has, again, its own scale of language endangerment which also indicates the certainty of the assessment, an element that is desirable for all such assessments.

The LinguistList itself holds a number of relevant resources, in particular MultiTree (LinguistList 2014), which collects genealogical trees and thus provides genealogical and further information on many languages and dialects, and LL-Map (LinguistList, n.d.), a similar collection of published language maps which links individual languages to the MultiTree resource.

This overview can only offer a selection of the most prominent work done in the field of language inventorying. More and more initiatives emerge, such as perhaps most recently the Wikitongues project (Wikitongues), similar in spirit to the Endangered Languages Project, more inclusive (not only endangered languages) but with less ties to the academic community.

Overall, we have seen much progress over the last few decades in our efforts to catalogue and map the languages of the world. We now can say with much confidence that there are around 6000 living languages in world if we apply mainly the linguistic criterion in a way so that varieties between which no or low mutual intelligibility exists are kept apart as different languages even in a dialect chain situation; and we know of some 500 more which are not spoken any more. If we allow for socio-political and functional criteria, we can identify at least around 7500 languages, and we can list them and refer to them with unambiguous codes.

As far as we can see at this point, what is needed is: (a) a movement to unite or combine the existing language catalogs with a quality check, for instance by establishing a recognized clearing house or permanent expert panel with an accepted academically recognized procedures to evaluate the status of individual languages; (b) more reliable language maps which better represent the complex and often multilingual language landscapes than dots or mostly non-overlapping colored areas can; (c) a more sustainable long-term institutional backing and funding so that the valuable resources created are reliably available in the future.

References

- Adelung, Johann Christoph & Johann Severin Vater. 1806. *Mithridates oder allgemeine Sprachenkunde – mit dem Vater Unser als Sprachprobe in bey nahe fünf hundert Sprachen u. Mundarten*. Reprint: Hildesheim: Olms, 1970. Berlin: Voss.
- Alliance for Linguistic Diversity. 2013ff. Endangered Languages Project. (<http://www.endangeredlanguages.com/>) (Accessed 1 May, 2018)
- Caines, Andrew & Christian Bentz. Languages of the world. (<https://cainesap.shinyapps.io/langmap/>) (Accessed 1 May, 2018)
- Caines, Andrew, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe & Paula Buttery. 2016. The Glottolog Data Explorer: Mapping the world's languages. Paper presented at the *VisLR II Workshop*.
- Dalby, David, David Barrett & Michael Mann. 1999. *The linguasphere register of the world's languages and speech communities. 1st ed. 2 vols.* Hebron, Wales, UK: Published for Observatoire Linguistique by Linguasphere Press / Gwasg y Byd Iaithe.
- Dixon, R. M. W. 2012. *Basic linguistic theory volume 3: Further grammatical topics*. Oxford, New York: Oxford University Press.
- Dobrin, Lise M., Jeff Good, William L. Svelmoe, Courtney Handman, Patience Epps, Herb Ladley & Kenneth S. Olson. 2009. SIL International and the disciplinary culture of linguistics. (Ed.) Lise M. Dobrin. *Language* 85(3). 618–658. (doi:10.1353/lan.0.0132)
- Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures (WALS) Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (<http://wals.info/>) (Accessed 1 May, 2018)
- Frawley, William (ed.). 2003. *International encyclopedia of linguistics, Second edition*. Oxford, New York: Oxford University Press.
- Gesner, Conrad. 1555. *Mithridates. De differentiis linguarum tum veterum tum quae hodie apud diversas nationes in toto orbe terrarum in usu sunt observationes*. Neudruck: Aalen: Scientia, 1974, hrsg. u. eingel. v. Manfred Peters. Zürich: Froschoverus.
- Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of language documentation*. Berlin, New York: Mouton, De Gruyter.
- Grimes, Barbara F. (ed.). 1984. *Ethnologue, 10th ed.* Dallas, Texas: Summer Institute of Linguistics.
- Grimes, Barbara F. (ed.). 1996. *Ethnologue, 13th ed.* Dallas, Texas: Summer Institute of Linguistics. (<http://www.ethnologue.com/13/>) (Accessed on 30 April, 2018)
- Grimes, Barbara F. 1998. Language names (by code). (<http://xml.coverpages.org/langcodesEth.txt>) (Accessed 30 April, 2018)
- Hale, Kenneth, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayeva Jeanne & Nora C. England. 1992. Endangered languages: On endangered languages and the safeguarding of diversity. *Language* 68. 1–42.
- Hammarström, Harald. 2005. Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics* 15. 34–45.
- Hammarström, Harald. 2015. Ethnologue 16/17/18th editions: A comprehensive review. *Language* 91(3). 723–737.
- Hammarström, Harald. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution* 1(1). 19–29. (doi:10.1093/jole/lzw002)

- Hammarström, Harald. 2018. Handbook of Descriptive Language Knowledge. (https://www.academia.edu/3142333/Handbook_of_Descriptive_Language_Knowledge) (Accessed 2 May, 2018)
- Hammarström, Harald, Sebastian Bank, Robert Forkel & Martin Haspelmath (eds.). 2018. Glottolog 3.2. Jena: Max Planck Institute for the Science of Human History. (<http://glottolog.org/>) (Accessed 1 May, 2018)
- Hammarström, Harald, Robert Forkel, Thom Castermans, Bettina Speckmann, Kevin Verbeek & Michel A. Westenberg. In prep. Visualizing language endangerment and language description hand in hand.
- Haspelmath, Martin. 2013. Can language identity be standardized? On Morey et al.'s critique of ISO 639-3. *Billet*. Diversity Linguistics Comment. (<https://dlc.hypotheses.org/610>) (Accessed 1 May, 2018)
- Haspelmath, Martin (ed.). 2014ff. Studies in Diversity Linguistics (book series). Berlin: Language Science Press. (<http://langsci-press.org/catalog/series/sidl>) (Accessed 26 April, 2018)
- Hervás y Panduro, Lorenzo. 1800. *Catalogo de las lenguas de las naciones conocidas, y numeracion, division y clase de estas segun la diversidad de sus idiomas y dialectos*. re-published as facsimile by the Biblioteca Nacional (España), Madrid, 2008.6 vols. Madrid: [s.ed.].
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195.
- Hockett, Charles Francis. 1958. *A course in modern linguistics*. New York: Macmillan.
- ISO 639-3. ISO 639-3. (<https://iso639-3.sil.org/>) (Accessed 1 May, 2018)
- Kaufman, Terrence. 1990. Language history in South America: What we know and how to know more. In Doris L. Payne (ed.), *Amazonian Linguistics: Studies in Lowland South American languages (Texas Linguistic Series)*, 13–73. Austin.
- Kaufman, Terrence. 2007. The native languages of South America. In R. E. Asher & Christopher Moseley (eds.), *Atlas of the world's languages*, 46–76. London: Routledge.
- Kloss, Heinz. 1967. Abstand Languages and Ausbau Languages. *Anthropological Linguistics* 9(7). 29–41.
- Krauss, Michael. 1992. The world's languages in crisis. *Language* 68. 4–10.
- Lewis, M. Paul & Gary F. Simons. 2010. Assessing endangerment: expanding Fishman's GIDS. *Revue roumaine de linguistique* 55(2). 103–120.
- LinguistList. 2014. MultiTree. A digital library of language relationships. Bloomington: Indiana University. (<http://multitree.org/>) (Accessed 1 May, 2018)
- LinguistList. n.d. LLMap: LL-MAP: Language and Location – A Map Annotation Project. (<http://www.llmap.org>) (Accessed 1 May, 2018)
- Meillet, A. & Marcel Cohen (eds.). 1952. *Les langues du monde: Par un groupe de linguistes, 2. Ed.*. Paris.
- Morey, Stephen, Mark W. Post & Victor A. Friedman. 2013. *The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization*. Handout of a talk given at the PARDISEC RRR Conference, Sidney. (<https://www.academia.edu/4725536/>) (Accessed 20 January, 2019)
- Moseley, Christopher (ed.). 2010. *UNESCO atlas of the world's languages in danger. 3rd ed.* Paris: UNESCO. (<http://www.unesco.org/languages-atlas/>) (Accessed 1 May, 2018)
- Moseley, Christopher & R. E. Asher (eds.). 1994. *Atlas of the world's languages*. New York: Routledge.

- Moseley, Christopher & R. E. Asher (eds.). 2007. *Atlas of the world's languages, 2nd ed.* New York: Routledge.
- MPI-EVA. 2015. Diversity Linguistics: Retrospect and Prospect | Presentations. (<https://www.eva.mpg.de/linguistics/conferences/diversity-linguistics-retrospect-and-prospect/presentations.html>) (Accessed 26 April, 2018)
- OLAC. 2003ff. Open Language Archives Community. (<http://www.language-archives.org/>)
- Pittman, Richard S. (ed.). 1951. *Ethnologue: languages of the world, 1st ed.* Dallas, Texas: Wycliffe Bible Translators (WBT). (<http://www.ethnologue.com/13/>) (Accessed 30 April, 2018)
- Ruhlen, Merrit. 1987. *A guide to the world's languages.* Stanford: Stanford University Press.
- Sebeok, Thomas (ed.). 1977. *Native languages of the Americas, volume 2.* New York, Boston: Springer.
- Skirgård, Hedvig. 2016. Clarifying points on Ethnologue pay-wall and language codes. Humans Who Read Grammars blog. (<http://humans-who-read-grammars.blogspot.com/2016/01/clarifying-points-on-ethnologue-pay.html>) (Accessed 1 May, 2018)
- UNESCO Ad hoc Expert Group on Endangered Languages. 2003. Language Vitality and Endangerment. UNESCO Document, presented at a meeting in Paris, 2003. (http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/Language_vitality_and_endangerment_EN.pdf) (Accessed 20 January, 2019)
- Vajda, Edward J. 2001. The Linguasphere Register of the World's Languages and Speech Communities (review). *Language* 77(3). 606–608. (doi:10.1353/lan.2001.0197)
- Voegelin, Charles F. & F. M. Voegelin. 1977. *Classification and index of the world's languages.* New York, Oxford, Amsterdam: Elsevier.
- WikiProject Languages. 2015. Primary language names in Ethnologue 13. Wikipedia. (https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Languages/Primary_language_names_in_Ethnologue_13&oldid=645050764) (Accessed 30 April, 2018)
- Wikitongues. (<https://wikitongues.org/>) (Accessed 1 May, 2018)
- Wurm, Stephen A. (ed.). 1996. *Atlas of the world's languages in danger of disappearing* (Theo Baumann, cartographer). Paris & Canberra: UNESCO.
- Wurm, Stephen Adolphe & Shiro Hattori. 1981. *Language Atlas of the Pacific Area: New Guinea area, Oceania, Australia; 2, Japan area, Taiwan-Formosa, Philippines, Mainland and insular South-East Asia.* Australian academy of the humanities.

Sebastian Drude
sdrude@hi.is

 orcid.org/0000-0002-2970-7996

Reflections on the diversity of participation in language documentation

I Wayan Arka
Australian National University
Udayana University

In this paper, I reflect on the diversity of participation in language documentation in the Indonesian context over the past two decades. I show that progress has been made in documentation research on the minority languages, with the concerted efforts of different stakeholders (community/non-community—among the latter, affiliations with universities, non-governmental organizations, the government, and other types of organizations of local speech communities). However, challenging issues remain in relation to the local communities' capacity, motivation, and leadership for helpful and long-term active participation in language documentation.

1. Introduction On this twentieth anniversary of Nikolaus Himmelmann's (1998) seminal article "Documentary and Descriptive Linguistics," I reflect on the diversity of participation in language documentation, mainly based on my experience in the Indonesian context.¹ Over the past two decades, much language documentation has been driven by the urgency of documenting endangered (typically minority) languages. Therefore, I begin my reflections by examining the participation of these targeted speech communities in the documentation process and discussing their roles and the extent of their contributions. I then reflect on the nature of the contributions of other stakeholders, such as the academic community, governmental institutions, and non-governmental organizations (NGOs). Finally, I summarize the issues raised and provide my personal assessment of the prospect of improving the participation of local communities in the Indonesian context and beyond.

¹I thank two anonymous reviewers for their helpful comments and feedback. I am also grateful to the editors for their invitation to contribute to this special volume in commemoration of the twentieth anniversary of Nikolaus Himmelmann's seminal article *Documentary and Descriptive Linguistics*.

2. Minority languages and participation of the speech community By minority languages, I mean the languages spoken by relatively small speech communities. The term “minority” is a relative notion, defined in terms of size and (in)equality in power and opportunities compared with the more dominant groups in a given geographical space. Speech communities with less than 1,000 members can definitely be considered minority groups in Indonesia. Based on this definition, 188 local minority languages exist in Indonesia (i.e., 34% of the total number of languages) (Arka 2013). In the ensuing discussion, I reflect largely on the participation of local minority communities in the documentation of their languages, which are either highly endangered, such as Marori in Merauke, Papua (119 people, with a dozen fluent speakers), or increasingly marginalized (though not yet endangered), such as Rongga in Flores (Arka 2010, 2015). Most, if not all, of Indonesia’s 188 minority languages would be considered endangered (cf. Anderbeck 2015).

The participation of local speech communities in documenting their languages is essential in any context. It can range from a simple role, such as giving permission (e.g., by a clan leader), to more complex tasks, such as participation in recorded speech events and other activities requiring specific or expert knowledge and skills (e.g., doing transcription using certain software tools, such as ELAN²). The participation can be light and casual, but it can also be intense and active. In most cases in the Indonesian context, community members tend to avoid intense involvement in language documentation and maintenance programs because such active participation often requires a high degree of motivation and a relatively high level of education, skills, and literacy to carry out documentation tasks. In modern language documentation, data collection and processing (transcription, helping with metadata, etc.) require community members’ literacy in using digital tools, such as video recorders and laptop computers. Furthermore, practical-educational work (e.g., developing learning materials for local schools) calls for basic knowledge of pedagogy and curriculum design. Seeking ongoing financial support for documentation projects demands skills in project proposal writing and access to possible funding networks at all levels, from local, regional, and national to international. In the contexts where I have worked, such requirements are too stringent to be met by local minority community members.^{3,4}

The primary concerns of many speech community members worldwide often involve meeting their basic needs—how to survive on a daily basis (e.g., food, housing, day-to-day finances, and jobs)—rather than the fate of their native languages. When day-to-day survival is a pressing concern, the motivation to participate in language documentation and maintenance can thus be low; nonetheless, many community members in such situations are happy to participate if paid to do so. The issue of long-term financial

²<https://tla.mpi.nl/tools/tla-tools/elan/>

³However, a local can acquire such qualities, typically after a process of capacity building and mentoring, provided that researchers are fortunate enough to have at least one capable local. For my Endangered Languages Documentation Programme (ELDP)-funded documentation project (<http://meraukelanguages.org/>), I have been fortunate to be assisted by Mr. Agus Mahuze, a highly motivated and capable young Marori, who happens to have a university education. With training and constant help, he has managed to acquire skills in data collection, data processing, and more importantly, develop further networks of his own for language documentation and language advocacy in Merauke.

⁴However, the reviewers point out that this is a highly context-dependent issue. In the North American context, many community language programs are directed by members of local speech communities, with and without academic training as linguists. Funding agencies have increasingly adapted their processes to facilitate funding provided directly to community-based organizations rather than academic researchers (cf. First Peoples’ Cultural Council. <http://www.fpcc.ca/email/email02201802.aspx>).

support for any project concerning language documentation and maintenance of minority languages remains a concern. The issue of motivation, particularly motivated leadership, can be one of the most difficult aspects of any community-based documentation project. For such a documentation enterprise to be successful and sustainable, at least one highly motivated local leader (if not a group of leaders) in the community must be willing to dedicate one's time and effort to documentation activities. I reflect more on the motivation issue in the final section of this paper.

3. Participation of the academic community As conceived by Himmelmann (1998), modern language documentation is essentially an academic enterprise in the field of linguistics. Unsurprisingly, most contributions to the growth of language documentation over the last two decades have come from the academic community, particularly linguists. These academics often collaborate with groups from other disciplines, such as ethnobiologists, ethnomusicologists, biologists, and computer scientists. All eighteen language documentation projects in Indonesia, funded by the Endangered Languages Documentation Programme (ELDP)⁵ and the Documentation of Endangered Languages (DOBES),⁶ over the past two decades have been undertaken by academics and students of linguistics.

Academics—either individually or collectively through their institutions and professional societies—have played pivotal roles in advancing the rapidly developing field of language documentation. The field has benefited greatly from the participation and contributions of different experts, including descriptive-typological linguists, sociolinguists, computer scientists, anthropologists, biologists, and ethnobiologists, as well as from the expertise of non-academic contributors, within and beyond the community. Linguistic programs now offer courses and even degrees in language documentation. Academics have also established publication outlets that are specifically intended for language documentation, for example, Language Documentation and Conservation (LDC)⁷ and Language Documentation and Description (LDD)⁸. Professional organizations, such as the Linguistic Society of America, initiate activities and provide platforms for scholars involved in documenting and archiving endangered languages to discuss and share solutions and intellectual advances in this new field (e.g., as documented in Grenoble and Furbee 2010). Conferences and workshops on language documentation are regularly organized (e.g., the bi-annual International Conference on Language Documentation and Conservation at the University of Hawaii), where commonly discussed academic topics include the creation and the use of various archival corpora, strategies for language maintenance/revival, interdisciplinary approaches to language documentation, technological advances in developing and using corpora, and promising directions for collaborative research. The above-mentioned concerted efforts of academics across different fronts have made language documentation a fast-developing field over the past two decades.

Academics are also the most active groups directly working with local communities in documenting their languages; equally important, they provide support for these activities. Based on my experience in documenting endangered languages in the Indonesian context, such support requires expertise in fields beyond linguistics because real issues in the field are multidimensional and often rooted in and mixed with complex

⁵<http://www.eldp.net>

⁶<http://dobes.mpi.nl>

⁷<http://nflrc.hawaii.edu/lcdc/>

⁸<http://www.elpublishing.org/publications>

socioecological problems (Arka 2005, 2008, 2013). These problems include difficulty in obtaining permission, power struggles and rivalry among community members or clans, and challenges encountered in convincing locals that language documentation as part of language maintenance is worth doing. Many academics involved in language documentation have been impacted by working closely with communities and have developed an understanding of the necessity of collaborative work within and beyond academic disciplines to respond adequately to the forces that create language shifts and endangerment.

Examining the affiliation of academic participation in language documentation in Indonesia shows an imbalance—most modern language documentation projects have been funded and led by foreigners. This situation is mainly due to funding bodies being based in developed countries, such as the United Kingdom and Germany, and grants being highly competitive internationally. Few Indonesian academics have applied for grants, and even fewer have managed to win them. This issue highlights the need for capacity building at both national and local levels in Indonesia. The increased diversity of practitioners of language documentation will guard against ethnocentric/Eurocentric myopias or biases and lead to better overall documentation of all languages.

4. Participation of governmental institutions Governmental participation in minority language documentation is often part of larger initiatives mandated by constitutions, legal acts, or charters. In this section, I reflect on the case of the Indonesian government's participation in the context of language policy and language management. Indonesia's 1945 constitution stipulates that local languages and cultures should be respected and maintained. The Department of Education and Cultures has therefore issued a series of ministerial decrees, resulting in the formation of what is now known as *Badan Pengembangan dan Pembinaan Bahasa* (National Board for Language Development and Cultivation) or *Badan Bahasa* for short.⁹

However, issues related to local languages are politically sensitive in Indonesia and dealt with as part of the broader national strategic language planning or language management efforts (Arka 2013; Moeliono 1994; Spolsky 2009). While *Badan Bahasa* assumes some responsibility for the documentation of local languages, as mandated by the 1945 constitution, a great deal of its time, effort, and resources has been devoted to the research on and the development of Bahasa Indonesia as a modern unifying language. The government's excessive emphasis on the unifying function of the national language, especially under President Suharto's regime from the 1970s to the 1990s, has caused negative unintended consequences for minority local languages across Indonesia.

An investigation on the publications of *Badan Bahasa* reveals the relative lack of attention to local languages (Arka 2013). For example, during the 1975–2007 period, 1,556 works were published, with only a third relating to local vernacular languages. Most publications focused on the healthy major languages, with Javanese topping the list (14.3%). The minority languages of eastern Indonesia received the least attention, with those in West Papua and Maluku having a 1:1 ratio (i.e., one language with one publication) or no publication at all. Many of the languages in this region are under-documented or undocumented and in urgent need of documentation.

After the fall of Suharto, Indonesia emerged as one of the most democratic countries in Asia. Local governments were granted greater autonomy, and West Papua was

⁹For the history of the different names of Badan Bahasa, see <http://badanbahasa.kemdikbud.go.id/lamanbahasa/sejarah>.

granted special autonomy status. Importantly, local languages—including teaching them as part of the *Muatan Lokal* (MULOK, meaning local-content curriculum)—are now the responsibilities of the local governments, as stipulated by the autonomy laws (Law 22/1999 on Local Autonomy and 2001 on Special Autonomy) and the law on languages (Law 24/2009). However, because of the politics of language in Indonesia, the ongoing tensions between the central and the local governments over the control of resources, as well as the lack of local capacity to handle the specialized tasks involved in dealing with local languages, I expect neither a radical change in the attention paid to minority local languages nor a significant increase in the resources devoted to them, especially in eastern Indonesia (see Arka 2013 for details). Little has changed in this respect since Indonesia embarked on its “big bang” decentralization or regional autonomy in 2001 although there is evidence that local autonomy has worked in certain other respects (see Hill 2014 for details).

My field experience suggests that while local autonomy may bring more freedom for locals to manage their own affairs, groups that are minorities in their own regions are still disadvantaged because the resources and the local language policies are controlled by the locally dominant groups. For example, in collaboration with the local Rongga teacher, I developed the MULO teaching material as part of my ELDP-funded documentation of Rongga (2004–2006). However, this teacher had an issue with using the material in class because of the policy that Rongga students must be taught the dominant Manggarai language as the MULO. As the curriculum and the timetable were already crowded, no time slot was available for teaching Rongga. Similar issues have arisen elsewhere in Indonesia. For example, in my documentation work with the minority Marori people in Merauke, Indonesian Papua, I observed that such inequality disadvantaged this group and violated the UN’s Universal Declaration of Linguistic Rights. Such inequality raises the need for external or higher-level governmental intervention and regulation to ensure that minority groups have access to the range of resources and support that they need to maintain their languages and cultures.

5. Participation of non-governmental organizations (NGOs) A number of NGOs are engaged in language and cultural documentation as part of their missions and activities. A well-known one is the Summer Institute of Linguistics (SIL International),¹⁰ which operates in the areas of research, training, and language material development. In Indonesia, the SIL has produced 1,383 publications (at the time of this writing), whose five leading domains covered are linguistic documentation (32.9%), language assessments (9.8%), anthropology (8.6%), sociolinguistics (5.6%), and literacy and education (3.3%).¹¹ The SIL has also conducted training workshops on language documentation in collaboration with universities and the government, particularly the Department of Education.

Historically, the SIL’s work in Indonesia has been related to the translation of the Bible as part of its Christian mission (Aritonang and Steenbrink 2008). The mixing of linguistic documentation with Christian missionary work has raised concerns in parts of Moslem-dominated Indonesia in recent years. These issues have resulted in the SIL’s difficulty in extending its permit to operate in Indonesia. Although it has been forced to close many of its branch offices across the country, the SIL staff members have continued their language

¹⁰<https://www.sil.org/about>

¹¹<https://www.sil.org/resources/search/country/Indonesia>. Accessed April 22, 2018.

documentation work in parts of Indonesia, including (West) Papua. The World Wide Fund for Nature (WWF) is another NGO engaged in documentation work in Indonesia. Its function that is related to local literacy and language documentation is part of its broader mission to maintain biodiversity and human ecology. My ELDP-funded ethnobiological documentation project on the Marori in Merauke (<http://meraukelanguages.org>) was carried out in collaboration with the local WWF.

Additionally, some smaller NGOs focus on individual local languages, for example, BASAbali (<http://basabali.org/?lang=bl>) in Bali. Its innovative approach to language conservation has attracted international attention, and it won the International Linguapax Award in 2018, in conjunction with International Mother Language Day. While targeting the Balinese language, BASAbali was in fact neither initiated nor directed by a Balinese. Nonetheless, its activities involve local language activists. BASAbali is “a collaboration of linguists, anthropologists, students and laypeople, from within and outside of Bali, who are collaborating to keep Balinese strong and sustainable.” International collaboration on this local project has proven to be successful, providing a fruitful model for the conservation of other local languages in Indonesia and beyond.

6. Final remarks: challenges and prospects In this final section, I reflect a little more on the contributions of local communities. If the success of language documentation is measured by not only the amount of rich, multipurpose, and multimedia datasets collected and processed but also its local impact, namely, the extent to which such documentation triggers the community’s awareness of its significance (e.g., for language maintenance), then considerable time and effort should be devoted to increasing the active participation of the local community. Given these goals, the major challenge of language documentation lies in motivating local people to participate in it. In my field experience, this is the most difficult aspect of the documentation process as it requires expertise and skills beyond linguistics. In particular, it calls for a deep understanding of local cultures and often, local politics. Such understandings are not quickly and easily acquired by an outsider. Hence, ideally, the best people to lead documentation projects are members of the community.

While it is desirable to engage the participation of locals, it is important that this involvement be useful. In Merauke, I found that more than enough people wanted to participate in my project, mainly to receive some cash in return. However, their contributions were not useful in most cases. This matter raises the issue of the local economy and capacity in carrying out language documentation. A discussion on local socioeconomic issues in language documentation is beyond the scope of this short paper, but I briefly explain capacity here.

Becoming a contributive participant (i.e., possessing the knowledge and the skills to process data, being able to provide data as required, or being able to organize locals) demands a certain degree of capacity, which can only be developed through training and education. Thus, capacity building should be an important component of any documentation project. Considerable effort, time, and funding should be devoted to capacity building. Ideally, this process should involve more than simply training participants in how to use modern technology for language documentation; it should also cover leadership and management training, including how to take advantage of external resources. For example, it is important to have a competent local leader or group of leaders who are able to write an external funding proposal. While I am aware that in some contexts, community-based and community-managed language documentation projects

are common, in the settings where I have worked, very few members of minority language groups have the capacity to be active and contributive participant-leaders.


Overall, my prognosis is that the road ahead for the language documentation enterprise, especially pursuing the goal of enlisting active and contributive participation of local community members, will not be easy. I expect that it will remain a challenge for language documentation practitioners for years to come, at least in Indonesia. The reason is that such participation results from locals being internally motivated, both individually and collectively. This motivation, combined with capacity, plays a critical role in determining how local stakeholders, as both individuals and groups, take up the modern challenges of language attrition and endangerment and how they respond strategically to these challenges. It is precisely this link between motivation and capacity, on one hand, and strategic action, on the other hand, that is so difficult to create at the local level. Both motivation and capacity are complex, interrelated processes, which involve cognition and local culture filters that are not always accessible to non-local researchers who want to help the local stakeholders. As pointed out in my earlier article Arka (2013), the roles of cognitive and cultural filters in language maintenance and revival have been overlooked in the literature. Therefore, one way to move forward is to pay more attention to these filters by integrating them into community-based programs. A logical and fruitful step would be to recruit anthropologists and educational psychologists as part of teams working with the community on language documentation projects. This initiative only addresses one of the problems (i.e., the motivation/participation issue). The issue of foreigners leading minority language documentation projects in Indonesia (and beyond) will unfortunately remain a difficult problem to tackle.

References

- Anderbeck, Karl. 2015. Portraits of Language Vitality in the Languages of Indonesia. In I Wayan Arka, Ni L Made Seri Malini & Ida A Made Puspani (eds.), *Language Documentation and Cultural Practices in the Austronesian World: Papers from 12-ICAL, Volume 4*, 19–47. Canberra: Asia-Pacific Linguistics.
- Aritonang, Jan Sihar & Karel Steenbrink. 2008. *A History of Christianity in Indonesia*. Leiden: Brill.
- Arka, I Wayan. 2005. Challenges and Prospect of Maintaining Rongga: A Preliminary Ethnographic Report. In Ilana Mushin (ed.), *Proceedings of the 2004 Conference of the Australian Linguistics Society*, 1–19. <http://hdl.handle.net/2123/138>.
- Arka, I Wayan. 2007. Local Autonomy, Local Capacity Building and Support for Minority Languages: Field Experiences from Indonesian. In D. Victoria Rau & Margaret Florey (eds.), *Documenting and Revitalizing Austronesian Languages*, 66–92. Honolulu: University of Hawai'i Press. <http://hdl.handle.net/10125/1353>
- Arka, I Wayan. 2010. Ritual Dance and Song in Language Documentation: Vera in Rongga and the Struggle over Culture and Tradition in Modern Manggarai-Indonesia. In Margaret Florey (ed.), *Language Endangerment in the Austronesian World: Challenges and Responses*, 90–109. Oxford: Oxford University Press.
- Arka, I Wayan. 2013. Language Management and Minority Language Maintenance in (eastern) Indonesia: Strategic Issues. *Language Documentation & Conservation* 7. 74–105.
- Arka, I Wayan. 2015. *Bahasa Rongga: Deskripsi, Tipologi dan Teori*. Jakarta: Pusat Kajian Bahasa dan Budaya, Universitas Katolik Atma Jaya.
- Grenoble, Lenore & Louanna Furbee. 2010. *Language Documentation : practice and values*. Amsterdam: John Benjamins.
- Himmelman, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Hill, Hal. 2014. *Regional Dynamics in a Decentralised Indonesia*. Singapore: ISEAS Publishing.
- Moeliono, Anton M. 1994. Indonesian Language Development and Cultivation. In Abdullah Hassan (ed.), *Language Planning in Southeast Asia*, 195–213. Kuala Lumpur: Dewan Bahasa dan Pustaka, Ministry of Education.
- Spolsky, Bernard. 2009. *Language Management*. Cambridge: Cambridge University Press.

I Wayan Arka

wayan.arka@anu.edu.au

 orcid.org/0000-0002-2819-6186

Reflections on software and technology for language documentation

Alexandre Arkhipov
University of Hamburg
Lomonosov Moscow State University

Nick Thieberger
University of Melbourne

Technological developments in the last decades enabled an unprecedented growth in volumes and quality of collected language data. Emerging challenges include ensuring the longevity of the records, making them accessible and reusable for fellow researchers as well as for the speech communities. These records are robust research data on which verifiable claims can be based and on which future research can be built, and are the basis for revitalization of cultural practices, including language and music performance. Recording, storage and analysis technologies become more lightweight and portable, allowing language speakers to actively participate in documentation activities. This also results in growing needs for training and support, and thus more interaction and collaboration between linguists, developers and speakers. Both cutting-edge speech technologies and crowdsourcing methods can be effectively used to overcome bottlenecks between different stages of analysis. While the endeavour to develop a single all-purpose integrated workbench for documentary linguists may not be achievable, investing in robust open interchange formats that can be accessed and enriched by independent pieces of software seems more promising for the near future.

1. Introduction¹ A major factor in the rise of language documentation (LD) since Himmelmann (1998) has been the move to digital methods, allowing an increase in

¹We thank the editors of this volume for their encouragement, and two anonymous reviewers for their valuable comments. All errors and possible shortcomings are our own. The contribution by A. Arkhipov has been made in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies

recordings and in the amount transcribed, and providing for analyses more firmly based in citable data than was previously the case.

The landscape for computer-assisted linguistic processing has changed considerably over the past 20 years. Fieldwork techniques have dramatically improved, with lightweight video and audio recording equipment, and new tools for annotating, analyzing and archiving language data. Further, with the emphasis on archiving has come better data management and methods for delivery of language records back to the source communities, not something that was easily or commonly done with analog recordings. Making these recordings available exemplifies the responsibility of academic researchers to work collaboratively and to ensure reusability of their data. As materials go back to the source community so also are new materials being created by members of that community.

Technology allows distant collaborations, and also allows for research to have multiple outputs from well-structured primary and secondary data and annotations. Differential access to language records between researchers and language speakers has dramatically reduced in many places and is rapidly shrinking in the rest. Mobile devices are making it easier to create and disseminate records of performance in small languages and social media promotes interactions in local languages. Digital media can be made available in formats derived from higher resolution archival formats, with descriptions that provide contextual information describing what is in the media. Ideally, transcripts of the contents of the media are also available and allow users to locate targeted points within cultural records.

However, the use of technology in recording language performance is not, in itself, sufficient to ensure the quality of the recording, nor to ensure its longevity. Thus, for instance, the position of a microphone is crucial to ensuring a good-quality recording, as is the use of a windscreen in outdoor settings. Analog recordings made on magnetic tape are now nearing the end of their playable life so digitization of the existing legacy of language recordings is one of today's urgent priorities. The standards specified by the international community of sound archives² need to be understood by linguistic researchers and applied routinely. As for the written domain, there was initial delight in being able to create complex documents like grammars and dictionaries using word-processors, but it soon became apparent that the proprietary formats they used could lock data away unless converted to another format.

There needs to be regular training in the use of technology so that the basic principles and standards are known and can be followed even as particular tools become obsolete. An example of a widely adopted standard is the Leipzig Glossing Rules³ which have improved cross-linguistic interpretation of glossed texts. It should be stressed that adhering to such standards is not an onerous condition on research and can take as little as reading a brief document or attending a training workshop. On the other hand, it must be acknowledged that although basic principles are quite straightforward to master, the details of use of particular tools and interaction between tools in different setups are highly specific and can often be a source of frustration. Thus not only an effort is required from the LD practitioners to invest in learning, but considerable effort is also required from the developers to invest in harmonization of tools and making workflows more straightforward and robust.

of Sciences and Humanities. Thieberger is an ARC Future Fellow and a CI in the ARC Centre of Excellence for the Dynamics of Language.

²<https://www.iasa-web.org/tc04/audio-preservation>

³<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>

In what follows, we start with a 20 years' flashback (§2), then proceed to review several important developments in technologies and workflows since that time (§3), and conclude with some speculative remarks on the future of technologies in LD (§4).

2. Back in 1998 Looking back to 1998, perhaps the biggest changes are in the capture, use, and analysis of dynamic media. Storage was expensive and data transfer took a long time. The first USB drives appeared in 2000 and the first terabyte disk in 2007.⁴ For audio recording equipment there was a choice between cassette recorders, minidisc (1992–2013), or DAT (1987–2005). Digital video had been available since 1986, using digital cassette tapes.

There were no simple transcription tools: ELAN⁵ was not yet developed, Transcriber⁶ was first released in 1998. In the same year, a hardware transcriber—a cassette player equipped with a pedal to repeatedly playback a portion of the tape—was the most sophisticated transcription device that Arkhipov saw used in the team fieldtrip of the Moscow State University. SoundIndex was an early transcription tool produced by Michel Jacobson at LACITO (Michailovsky et al. 2014) and was used in the first online presentation of text and media (now PANGLOSS). SoundIndex was used by Thieberger in his analysis of Nafsan (South Efate) and allowed his grammar to be the first to cite examples back to an archival media corpus (Thieberger 2009).

In 1998 the only digital indigenous language archive was text-based (the Aboriginal Studies Electronic Data Archive – ASEDA; see Thieberger 1995). The Archive of the Indigenous Languages of Latin America (AILLA)⁷ began in 2000, the DoBeS programme in 2000 started the MPI Archive in Nijmegen (now the TLA).⁸ The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)⁹ began in 2003, and the Endangered Languages Archive (ELAR)¹⁰ in 2004. There are now a number of other such archives (see the Open Language Archives Community (OLAC)¹¹ page or DELAMAN¹² for lists of affiliated archives). Associated with digital archiving is the widespread adoption of the metadata standards established by OLAC and the TLA. These have allowed significant information sharing about the content of archives and increased access to the records they hold.

3. Advances and emerging challenges over the past 20 years In this section, we briefly review several deliberately selected technology-related issues, necessarily leaving aside many others, by no means less important. We will touch upon ensuring longevity of LD data (§3.1), as well as upon transitioning between various analysis stages (§3.2) and tools (§3.4), going portable (§3.3), publishing LD data (§3.5) and, finally, training in technologies (§3.6).

3.1 Data preservation and management While there are many benefits of digital technology, problems have become more apparent with experience. The fragility of

⁴<http://www.computerhistory.org/timeline/memory-storage>

⁵<https://tla.mpi.nl/tools/tla-tools/elan/>

⁶<http://trans.sourceforge.net/en/presentation.php>

⁷<https://ailla.utexas.org/>

⁸<https://tla.mpi.nl/>

⁹<http://paradisec.org.au>

¹⁰<https://www.soas.ac.uk/elar/>

¹¹<http://www.language-archives.org/>

¹²<http://delaman.org>

digital data means paying attention to backing up, especially in climates where equipment life is compromised by moisture. Being able to record more easily has the corollary of creating many more files to manage. Data (and metadata) management is now a serious consideration (in all research, not just in linguistics, see Corti et al. 2014) and is being taught as part of field methods courses.

Aside from the preservation and management of digital files themselves, the evolution of hardware and software requires continuous migration of data to newer formats and storage media. As software has come and gone over the past two decades it has prompted thinking about how our data can survive the tools we use. Many of us have had the experience of finding files created in the past that are no longer accessible. It is also enticing to produce a multimedia app or website that has rich content, but, again, the primary data used in these has to be kept in an open format, as multimedia products have a very short lifespan. Sustaining data means ensuring it can continue to be read and that there are copies of it in a number of locations, ideally also in an archive. To do this, it is best to keep an open format or text copy of files rather than have them stored inside a proprietary format (like Microsoft formats xls or doc).

A concomitant problem, very acute before the Unicode standard came into wide use, was competing and idiosyncratic character encodings and fonts (see also Kalish 2007 on this issue). Combining two or more character sets in a single document presented a particular challenge, such as writing a descriptive grammar in Russian with English glosses and custom transcriptions adorned with various diacritics. Years later, even if the file format is still readable by modern software, half of the characters are hard to reconstruct. Transcoding solutions such as SILConverters¹³ proved to be particularly helpful in such cases.

3.2 Transcription bottleneck With this increased volume of recordings, transcription has become a major bottleneck (see Himmelmann this volume). In fact, only a fraction of the collected data ever gets transcribed, which means that even smaller data volumes end up as fully analyzed corpora.

However, LD will be able to increasingly benefit from the technologies developed for major languages, both in spoken and written form. ASR (automatic speech recognition) and related speech technologies such as forced alignment have been shown to efficiently reduce the transcription-related manual workload. Although training acoustic models used in speech recognition normally requires large volumes (sometimes hundreds of hours) of annotated data, different methods are emerging to reduce the effort of porting such systems to work effectively on smaller speech corpora (see Strunk et al., 2014; Adams et al., 2018; Johnson et al., 2018).

Another option to overcome the bottleneck is delegating the work to native speakers. The Basic Oral Language Documentation framework (BOLD; see Reiman 2010) and similar approaches suggest recording careful re-speaking of the analyzed text by (the same or another) native speaker, which can much more easily be further transcribed by linguists on their own; other kinds of oral annotations such as oral translation or comments can also be provided. SayMore is currently a tool that supports recording both oral annotations and oral translations; it is reported to be successfully used in documenting 14 languages of Nepal (Khadgi 2017).

¹³<http://scripts.sil.org/encnvtrs>

Other initiatives support transcription crowdsourcing through an online repository where people can register to provide transcriptions for items of their choice. One such project is *Euskal Herriko Ahotsak* (Voices of the Basque Country),¹⁴ an archive of thematic interviews with speakers of diverse varieties of Basque. Another is Phonemica,¹⁵ a collection of stories in languages of China, where the recordings themselves are also contributed by the users and can be transcribed and translated online into Mandarin and English.

3.3 Portable solutions Another tendency of recent years, paralleling the growing performance and shrinking footprint of hardware, is the increasing demand for more lightweight and portable technological solutions. After desktop computers came laptops, by now ordinary and cherished companions of a fieldworker, then tablets and smartphones. Computers, formerly confined to pre- and post-fieldwork office use, are now indispensable throughout the field session. Portable devices and field conditions impose limitations on the software, including memory use and system performance, undesired dependencies on particular operating systems, frequent updates and connectivity. While tablets and smartphones may be unadapted to more complex analytical work, they can be a lifesaver when it comes to simpler operations which do not wait, like taking quick notes (including oral notes), collecting metadata or looking up a word in the dictionary. Higher quality devices can also be used to collect primary data, be it photos, video or audio—something which 20 years ago would require three separate and bulky analog devices.

For a long time, documenting a language was mostly seen as the linguist's domain. Nowadays, members of the speech community are not just 'contributing' to the documentation, but are frequently taking a leading role. Accordingly, the need arises to train the speakers to use linguistic equipment and software, or, more wisely, to produce tools which can easily be mastered by non-linguists. To name just a few, the dictionary collecting tool WeSay¹⁶ and the organizer-and-transcriber SayMore¹⁷ (both for Windows PCs), audio collecting and translating app LIG-Aikuma,¹⁸ and Zahwa¹⁹ app for documenting procedural knowledge like food cooking recipes (both for Android devices) have been successfully used to collect data in many remote locations across continents. Some of them are also integrated with bigger applications like FLEx²⁰ or ELAN, and/or offer options of preparing standardised archive submissions. Functions of oral annotations (see above) drastically lower the barrier of required user expertise.

3.4 Tool interoperability and data structures Language documentation comprises an array of diverse activities, each with its own focus and demands in data treatment. First linguistic tools that came into existence were rather specialized and each tackled a very limited portion of the workflow. For instance, the ability to transcribe directly into a digital format was a huge productivity boost by itself. However, while more and more data became produced and processed on different stages of the workflow, the transitions back and forth between transcripts, interlinear glosses, dictionaries became a new bottleneck.

¹⁴<https://ahotsak.eus/english/>

¹⁵<http://phonemica.net/>

¹⁶<https://software.sil.org/wesay/>

¹⁷<https://software.sil.org/saymore/>

¹⁸<https://lig-aikuma.imag.fr/> see also <http://www.aikuma.org/>

¹⁹<https://zahwa.aikuma.org/>

²⁰<https://software.sil.org/fieldworks/>

Early tools like Toolbox²¹ have determined the data structures used by linguists so that e.g. the common format for many bilingual dictionaries of small languages is the ‘backslash’ file (the SIL ‘Standard Format’).²² These files are plain text and can be converted to new formats and be archived, although there is considerable variation in user-specific field codes and structure. The same format was used for storing interlinear texts, with the additional problem that it relied on counting whitespace characters for word-by-word alignment.

An important step towards better interoperability and sustainability was the adoption of XML (introduced in 1998)²³ either as native format or at least as an export/import format by most tools. While data structures embodied in XML documents vary greatly between different software, the availability of common standard-compliant processing tools makes it technically possible and relatively easy to convert between them. Yet the multitude of tools as well as their quick evolution is a challenge. In an LD project that Arkhipov took part in from 2006, all software elements and data formats used at different steps for transcribing, glossing, archiving and presentation changed within 3-5 years, which demanded considerable effort to maintain.

In the currently running long-term INEL project,²⁴ transcriptions coming from four different sources are imported into FLEx for glossing: plain text typed in from archival manuscripts, transcripts by native speakers in common office format, transcripts done in SayMore by more computer-proficient native speakers and those made in ELAN by linguists. Once glossed, the texts are exported into EXMARaLDA²⁵ format for further annotation and presentation. This is all possible thanks to the interaction between ELAN and FLEx which improved greatly since 2008, now preserving speaker, time-alignment and mediafile attributes crucial for time-aligned glossed text corpora. However, the inability of FLEx to import existing morpheme glosses remains a major blocker. It not only makes it impossible to incorporate external changes to any aspect of at least partly glossed text, but also prevents many from using FLEx altogether, especially those having a substantial corpus analyzed elsewhere (e.g. in Toolbox or Word). Another one is lack of support for custom annotation tiers in FLEx. These two problems are however not inherent to the FLEx interlinear XML format, which curiously is sometimes used as a pivot interchange format without accessing the FLEx application itself.

Two alternate ways of dealing with interoperability problem can be distinguished. One is adding functionality to an existing tool, thereby reducing the need to interact with other tools. Thus ELAN as primarily a multimedia transcription/annotation tool now includes a glossing module, FLEx integrates lexicon and text analysis with dictionary-publishing solutions, and SayMore combines metadata curation and file management with transcription. However, it seems that the complexity of the most sophisticated tools is close to reaching a certain limit beyond which the required development and maintenance efforts surpass the resources of the linguistic community. A hypothetical all-purpose workbench for documentary linguists would need to support a wide range of primary data including various media types, various content types (texts, words, paradigms, questionnaires, as well as metadata), flexible annotations, interlinking between text and media, complex analytical tools, rich visualization and publishing options. Not only does

²¹<https://software.sil.org/toolbox/>

²²http://downloads.sil.org/legacy/shoebox/MDF_2000.pdf

²³<https://www.w3.org/TR/1998/REC-xml-19980210>

²⁴<https://inel.corpora.uni-hamburg.de/>

²⁵<http://exmaralda.org/en/>

such omni-functionality require deep insight into possible use cases to be adequately designed and an extremely diverse developer expertise, but the application becomes increasingly heavy and slow which ultimately restricts possible use cases (cf. §3.3). On the other hand, a universal tool must also be cross-platform: meanwhile, developer experience tells us that in this case a lion's share of worktime is taken by making each and every feature function identically in different computational environments. These are perhaps the most evident reasons why a single all-purpose LD tool is unlikely to appear in the foreseeable future.

An alternative is, then, to ensure lossless and efficient omnidirectional data transfer between independently developed tools. In this view, open, well-documented (and possibly human-readable) interchange formats are a priority. A recent release of CLDF, a specification for Cross-Linguistic Data Formats,²⁶ is a promising step along this path. It proposes an interchange standard for linguistic datasets representable in tabular form. Leveraging the positive experience of a series of cross-linguistic projects such as WALS and Glottolog,²⁷ it explicitly aims to decouple the development of software tools from that of datasets. CLDF makes use of plain text tab-delimited files, which can be read and edited by humans and on the other hand are supported by a wide range of software. This makes the format particularly suitable for configuring custom tool chains for multi-step data processing.

3.5 Data vs. presentation It is important to present language records in a way that is accessible and attractive, but this presentation should be considered as a kind of exhibition derived from the underlying collection, not as the only product of documentation. A typical example is a lexical database in which as much information as possible is stored about each word, ideally including encyclopaedic information. Dictionaries of various kinds can be derived from this lexical database: a detailed dictionary, a learner's dictionary, or a topical dictionary. These can be presented in several ways: on paper, as a website, as an app. While a book or website can go out of date or be lost, the primary data must continue to be available for use in future. Similarly, a digital corpus can be presented as a resource for linguists, with full annotation and complex search facilities, or as a text collection with a focus on the speakers and the story, for a wider audience including the speech community.

Putting aside large archives with established infrastructure, there is currently no widely accepted and easily reusable solution for publishing LD data in a user-friendly manner. For lexical data, one should mention the no longer developed LexiquePro²⁸ which can generate a set of static HTML pages (for publishing on one's own website), and the current online Webonary²⁹ repository which allows users to register and upload their Toolbox or FLEx lexical databases for general access, providing dynamic browsing and search capabilities (although little customization). Similar attempts have been made for sharing interlinear texts, such as the Kratylos³⁰ project (yet rather at proof-of-concept stage). More powerful solutions, such as the recent Tsakorpus³¹ corpus platform,

²⁶<http://clfd.clld.org/>

²⁷<http://glottolog.org/>

²⁸<http://www.lexiquepro.com/>

²⁹<https://www.webonary.org/>

³⁰<https://www.kratylos.org/>

³¹<https://bitbucket.org/tsakorpus/>

generally require installing and configuring one's own instance of a server application which assumes an available server and substantial technical competence.

3.6 Training and support As new tools and methods appear, there is a need to train LD practitioners and to provide advice about emerging devices (cameras, recorders, microphones and so on). Such training has been provided at summer schools or training workshops, such as those run by InField/CoLang,³² LLL,³³ ELDP,³⁴ or DoBeS.³⁵ An email list and website run by the Resource Network for Linguistic Diversity³⁶ has provided advice for subscribers since 2004. As software evolves even more rapidly than hardware, can be quite complex and contain poorly documented features and bugs, one-off training is usually not enough to ensure seamless work in the future. Contacting developers directly is not rarely the most efficient strategy, however not always realistic; one then has to rely on the user community and fellow researchers for continuous support and advice. Also the issue of translating the user interface and help pages should not be neglected. Here again, the LD user community is often an important actor, since developers' resources are limited.

4. Future of LD technologies Building appropriate methods into normal fieldwork practices results in records that can be archived and so made accessible to source communities. These records are robust research data on which verifiable claims can be based and on which future research can be built, and are the basis for revitalization of cultural practices, including language and music performance.

Looking ahead we can predict that recording and storage technologies will become cheaper, smaller and more intuitive, which will make it easier for more documentation by speakers, increasing the need for LD networks to train speakers and to provide ways of storing the records they create into the future. As a means of sharing these records, social networks and media platforms are likely to increasingly become centres of activity in writing, recording and distributing language performance. Crowdsourcing is starting to be used for transcribing, translating and annotating the collected data. At the same time, advanced technologies related to speech recognition, translation, text-based annotation will be increasingly applied to LD data as another remedy against the 'transcription bottleneck'.

On the other hand, harmonising, interlinking and reusing data across projects will require increased attention to develop and promote standardised software-independent formats for various data types. Such formats would allow independent tools and services to access a portion of data from a repository, process it and return enriched data which can then further be accessed by other tools and users. Provided a standard interface to access arbitrary small pieces of data, dynamic annotations similar to formulas in a spreadsheet document could become instrumental to speed up annotating large corpora and to enable updating interdependent annotations. Such architecture would also facilitate creating varied presentation formats from the same linguistic data, addressing different audiences and adapting to different devices and environments.

³²<https://www.alaska.edu/colang2016/>

³³http://www.ddl.cnrs.fr/colloques/31_2012/

³⁴<http://www.eldp.net/en/our+trainings/about/>

³⁵http://dobes.mpi.nl/dobesprogramme/training_courses/


³⁶<http://rnlld.org>

References

- Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In Calzolari, Nicoletta (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 3356–3365.
- Corti, Louise, Veerle van den Eynden, Libby Bishop & Matthew Woollard. 2014. *Managing and sharing research Data: A guide to good practice*. London: Sage Publications.
- Johnson, Lisa M., Marianna Di Paolo & Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation & Conservation* 12. 80–123. <http://hdl.handle.net/10125/24763>
- Kalish, Mia. 2007. Review of Fontographer. *Language Documentation & Conservation* 1(2): 301–311. <http://hdl.handle.net/10125/1723>
- Khadgi, Mari-Sisko. 2017. Large-scale language documentation in Nepal: A strategy based on SayMore and BOLD. Paper presented at 5th International Conference on Language Documentation and Conservation (ICLDC), Honolulu, Hawai‘i, March 4, 2017. <http://hdl.handle.net/10125/42029>
- Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, S everine Guillaume, Alexandre Fran ois, Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation & Conservation* 8: 119–135. <http://hdl.handle.net/10125/4621>
- Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4. 254–268. <http://hdl.handle.net/10125/4479>
- Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association (ELRA), 3940–3947. <http://www.lrec-conf.org/proceedings/lrec2014/summaries/1176.html>
- Thieberger, Nick. 1995. The Aboriginal Studies Electronic Data Archive, *International Journal on the Sociology of Language* 113. 147–150.
- Thieberger, Nick. 2006. Computers in Field Linguistics. In Keith Brown (ed.), *Encyclopedia of language & linguistics*, 2nd edn., vol. 2, 780–783. Oxford: Elsevier. <http://hdl.handle.net/11343/34940>

Thieberger, Nick. 2009. Steps toward a grammar embedded in data. In Patience Epps & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389–408. Berlin: Mouton de Gruyter. <http://repository.unimelb.edu.au/10187/4864>

Alexandre Arkhipov
alexandre.arkhipov@uni-hamburg.de

Nick Thieberger
thien@unimelb.edu.au
 orcid.org/0000-0001-8797-1018

**Beyond Description:
Creating & Utilizing
Documentations**

Reflections on descriptive and documentary adequacy

Sonja Riesberg

Universität zu Köln

CoEDL Australian National University

One of Himmelmann's primary goals in his 1998 paper was to argue for a strict division of documentation and description. Language documentation has since successfully developed to become a discipline in its own right. Nevertheless, the question concerning the interrelation of description (and thus analysis) and documentation remains a matter of controversy. This paper reflects on descriptive and documentary adequacy, focusing on two major issues. First, it addresses the question of how much analysis should enter into an adequate documentation of a language and, second, it discusses the role of language documentation and primary data in the replicability of linguistic analyses.

I started my linguistic career as a PhD student in a documentation project funded within the Volkswagen Foundation's DoBeS program. The departments I have been affiliated with since have had a strong focus on documentary linguistics and the great majority of my colleagues and friends there are documentary linguists. Thus, to me, reading Himmelmann 1998 feels almost outdated; a statement of the obvious. In the 20 years since its appearance, the field has changed dramatically. A huge number of documentation projects have been funded, and a significant part of the linguistic community naturally considers language documentation to be one of the many linguistic disciplines. Himmelmann 1998 quite obviously had an immense impact. The fact that it is still shockingly relevant was something that I only slowly started to realize when I emerged from my PhD bubble and came to see a greater variety in how people 'do linguistics'. Reflecting on the topics as they were originally raised in the paper, how they have been received and how they have developed over the years is thus probably not as redundant as it might seem at first sight.¹

¹I am grateful to Birgit Hellwig, Stefan Schnell, and Sarah Verlage for having reflected with me on the one or other thought addressed in this short paper. Thanks also to two anonymous reviewers and the editors of this volume, especially Bradley McDonnell, for valuable comments.

The delimitation of documentation and description is one of the core issues discussed in Himmelmann 1998, and one of the explicit goals of the paper was that “the collection and presentation of primary data [should] receive the theoretical and practical attention they deserve” (Himmelmann 1998: 164). Interestingly, the demand for a strict separation of documentary and descriptive activities has, at times, caused rather emotional reactions in the linguistic community, and still does today. Partly, I believe, this is due to a misunderstanding and misconception of the paper. Some of the quotes one finds about Himmelmann 1998 suggest that the paper was not read in the first place, or maybe at best skimmed, in order to take quotes out of context and thereby miss the point entirely. Anyone who insists that Himmelmann advocates data collection without analysis has obviously not read (or understood) the paper. Anyone who claims that for Himmelmann language documentation is (nothing but) “a radically expanded text collection” has started reading, but quite obviously stopped halfway through. One very recent example is the introductory chapter to the *Oxford Handbook of Endangered Languages*. In an overview that is to answer the question “what is language documentation,” Campbell & Rehg list a few quotes, such as “a language documentation is a lasting, multipurpose record of a language” (Woodbury 2011: 159), the Hans Rausing Endangered Languages Project website, and the above-mentioned Himmelmann quote about the “expanded text collection” (Himmelmann 1998: 165). They then conclude that “with statements such as these, it is little wonder that some linguists [...] have misinterpreted this approach to mean: Documentary linguistics is all about technology and (digital) archiving. Documentary linguistics is just concerned with (mindlessly) collecting heaps of data without any concern for analysis and structure. Documentary linguistics is actually opposed to analysis” (Campbell & Rehg 2018: 11). I disagree. I don’t believe it is the statements themselves that trigger this misinterpretation, because typically they are elaborated on in more detail in the original sources, which rarely leave room for misunderstandings. It is the way some authors choose these statements and list them in isolation, that—intentionally or not—promote a skewed picture.

In any case, it seems that an interpretation of Himmelmann has developed and continues to be spread that attributes to Himmelmann the idea that language documentation does not (or even must not) include analysis. This might be one reason for certain developments in the field, which in my opinion are rather unfortunate. One such idea is that documentation corpora should contain no elicited material at all; for instance, recordings of elicitation sessions where linguists and native speakers work on lexical, phonological/phonetic, or grammatical problems. There might, of course, be various other reasons for this, such as following the wishes of the speech community to include only ‘nice’ and ‘clean’ data in the collection, or the embarrassment felt by the researcher upon listening to themselves struggling to come to grips with complex grammatical issues, or trying to determine the exact semantics of a culturally specific lexical item. But I think the reluctance to archive elicitation data and make them available to the wider public partly also stems from the widespread idea that a language documentation is to include ‘natural’ data only, and that elicitation should not be part of it. To my thinking, this is an interpretation of Himmelmann’s demand to separate descriptive and documentary activities which clearly overshoots his goal. In this sense, I agree with Berge’s (2010) notions of adequacy in documentation, in that a documentation should include “basic phonology, morphology, syntactic constructions with context, lexicon, a full range of textual genres, registers, and dialects, and data from diverse situations and speakers” (58). It is the large variety of data, including natural data as well as elicited materials, that

makes a good documentary corpus such a valuable resource. Yet, another unfortunate trend can be observed when looking at the practices of some funding agencies that support documentation projects. Here, one sometimes gets the impression that, for the agencies, the important aspect of ‘value for money’ is only evaluated in terms of hours of (natural language) recordings plus transcriptions that are deposited in the archives. But this is not, of course, what we conceive of as a comprehensive and adequate language documentation.²

These reactions and developments are all the more surprising considering that Himmelmann 1998 explicitly addresses the close interrelation of language documentation and language description at different points, and clearly states that his approach “does not imply that it is possible to make a ‘pure’ documentation without any descriptive analysis” (Himmelmann 1998: 165). The point Himmelmann made was that primary data should not be collected solely for the purpose of description (and then kept in some drawer (until the linguist’s death) to be disposed of (after the linguist’s death)). Instead, description was seen as part of the documentation. That is, description is considered a necessary part of documentation, but not its purpose. In Himmelmann’s view “a good and comprehensive documentation will include all the information that may be found in a good and comprehensive descriptive grammar” (170). The opposite is not the case: The most detailed and carefully researched grammar can never cover what a language documentation covers, simply because it is a written medium that cannot compete with the multi-media resources that make up a language documentation. Exactly when description should therefore be postponed to save resources for good quality documentation, as suggested in a later paper by Himmelmann (2006: 24), is a matter of controversy (see, e.g., Evans 2008: 346).

The second reason why the paper has caused emotional reactions is because it challenges the way in which we as a discipline have been working and how we have treated primary data for decades. There are two general and substantial points to this issue. The first concerns the question of what kind of database we, as a field, want to ground our knowledge in. The second concerns the question of replicability. As Himmelmann states, “a language description aims at the record of A LANGUAGE, with ‘language’ being understood as a system of abstract elements, constructions, and rules that constitute the invariant underlying structure of the utterances observable in a speech community. A language documentation, on the other hand, aims at the record of THE LINGUISTIC PRACTICES AND TRADITIONS OF A SPEECH COMMUNITY” (Himmelmann 1998: 166). And, as mentioned above, the assumption is that a good documentation will include all the information that is also included in a good description. The important thing to note is that if I base my description on a comprehensive documentation, I will have to account for all the variation I find in this compilation of primary data. This is the ideal scenario, in which I will account for—or at least describe—all (morpho-)syntactic structures that occur in my corpus, all the inter- (and intra-)speaker variation, all the inter- (and intra-)genre variation, etc. If, however, I approach a speech community and their language with the sole purpose of grammar writing, it is likely that much of the above will not make it into my description, because my data base is in all likelihood a very different one. In the worst case it will consist of elicitations of phenomena I decided to investigate beforehand (e.g.,

²See also Himmelmann (2012) for a more detailed argumentation against the misconception that language documentation is equivalent to “(mindlessly) collecting heaps of data” and opposed to analysis (187), and, e.g., Caballero (2017) for a very nice report on how diverse a documentation corpus can be and how different data types in a corpus can be utilized by different stakeholders.

verbal and pronominal paradigms, clause chaining, relativization, etc.), elicited with one or two speakers only, and possibly illustrated with one or two carefully chosen examples from one or two narratives I collected, perhaps specifically for this reason. Obviously, the two scenarios depicted constitute the two endpoints of a continuum and most linguists will find themselves doing something in between. For instance, the ‘traditional’ field linguist writing a grammar will collect more than just one or two narratives and, of course, will also treat phenomena that go beyond such questions as one finds, e.g., in precast questionnaires. On the other hand, a single person working on a language documentation will not (in one lifetime, and even less so within the usual funding period) be able to describe everything there is in her collection. But the question of which approach will result in a more adequate record of the language investigated is easy to answer, and I am convinced that as a field we should strive towards the ideal, albeit in the knowledge that it represents more than a lifetime’s work. Yet, it is still the case that language documentations count less than the written word. Many documentary linguists, especially early career researchers, who will often have spent years of their career compiling language documentation corpora, will probably have had the painful experience that an appointment committee has asked the question why ‘so little’ has been published. This is not to say that the community is unaware of this problem. It is a widely discussed deficit, and there have been serious attempts to enhance the status of documentation corpora, to review them and thus make them utilizable for application processes. We are just not quite there yet.

Turning to the question of replicability, this involves making primary data accessible. It feels like there should not be much to say about this. If we believe that science can only be taken seriously if quality can be scrutinized, primary data should to be made available (if possible), and they should be made available in a format that allows for verification or falsification of the claims made. Needless to say, this is not achieved by simply uploading an audio or video file to a digital archive, or handing over a collection of cassettes or external hard drives to physical one.³ The falsifiability of descriptive statements was not one of the primary concerns in Himmelmann 1998 (see Himmelmann 2006: 15, where this issue is discussed in a little more detail), though it is mentioned in passing: “it is simply a feature of scientific enterprise to make one’s primary data accessible for scrutiny” (p.165). Compared to other disciplines, (typological) linguistics has a lot of catching up to do. In psychology, for example, journals strongly encourage their authors to “deposit research data in a relevant data repository and cite and link to this dataset in their articles”. This is considered the default, and if, for some reason, it is not possible to make data available, the author has to “make a statement explaining why research data cannot be shared”.⁴ In other fields data sharing is not only *encouraged* but even *required*. This is still pretty much unheard of for linguistic journals, though people have started discussing the topic of replicability and there is a serious demand for data sharing also in linguistics (see, e.g., the paper by Gawne & Berez-Kroeker, this volume). The utilization of documentation is probably one of the major foci of attention in the current discourse on documentary linguistics. Yet, until today, descriptive statements – especially about small, underdescribed languages – are hardly ever falsifiable. Extended appendices with

³This, of course, also holds if we want our data to be useful to researchers from other disciplines and to members of the respective speech communities.

⁴These two quotes originate from the authors’ information for the journal *Cognitive Development*, cf. <https://www.journals.elsevier.com/cognitive-development>. See also <https://www.elsevier.com/authors/author-services/research-data/data-guidelines>.

interlinearized and translated texts that are added at the end of a grammar give some opportunity for falsifiability. In short research papers with limited space this becomes more difficult, especially for topics that are not well captured through transcription, not to mention studies on tonal phenomena and prosody. Language documentation corpora are one possibility, and probably the most adequate and comprehensive one, to address this issue. This is, on the one hand, because modern language archives usually (try to) guarantee long-term preservation, and on the other hand, because language documentations offer the amount and variety of data necessary to make falsifiability possible in the first place. Surprisingly, many linguists who have spent a lot of time and energy compiling these corpora (including the immensely time-consuming transcriptions, annotations, and commentaries) do not make use of these invaluable resources and often do not even reference them in their publications, even if they form the basis of their research. Others (and, I'm afraid, I partly include myself here) will make a general statement, hidden in a footnote, along the lines of "all data used in this paper is available online under xxx," but then do not make the effort to link examples to the original source. A nice example of how this can be done successfully is Seifart et al.'s recent paper on Bora drummed speech (Seifart et al. 2018), which links to the whole Bora language documentation corpus (Seifart et al. 2009), but also to single, relevant sessions (i.e., video and audio recording plus transcriptions and translations in an ELAN file) that illustrate the statements made in the paper.

Obviously, a lot has been achieved in the field of language description and documentation since Himmelmann 1998 was published (so much in fact, that it took me until about four to five years into my academic career to realize that the practices sketched in this paper are not to be taken for granted). It is just as obvious, though, that there is room for improvement. This pertains to the role of language documentations in helping linguistics to catch up with other sciences in terms of replicability of research results. But it also pertains to Himmelmann's vision quoted at the beginning of these reflections that documentary activities should receive the attention they deserve. They do not, as long as a paper in any mid-range linguistic journal on a researcher's CV still means more to potential employers and funding agencies than an annotated multi-media corpus with careful commentary of an endangered language. As mentioned above, both these issues are currently avidly discussed in the documentary community. What seems to receive less attention is the rather paradoxical development that due to the very success of establishing language documentation as a discipline, it has lost its descriptive component, which was always supposed to be—and I believe should be—an important and substantial part of it.

References

- Berge, Anna. 2010. Adequacy in documentation. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language Documentation: Practices and values*, 51–66. Amsterdam: John Benjamins.
- Campbell, Lyle & Kenneth R. Rehg. 2018. Introduction: Endangered languages. In Kenneth L. Rehg & Lyle Campbell (eds.), *The Oxford handbook of endangered languages*, 1–20. Oxford University Press.
- Evans, Nicholas. 2008. Review of Essentials of language documentation. *Language Documentation & Conservation* 2(2). 340–350.
- Gawne, Lauren & Andrea L. Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation on the 20 year anniversary of Himmelmann 1998*. <http://hdl.handle.net/10125/24805>
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36, 161–195.
- Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds), *Essentials of language documentation*, 1–30. Berlin and New York: Mouton de Gruyter.
- Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation* 6. 187–207.
- Seifart, Frank, Julien Meyer, Sven Grawunder & Laure Dentel. 2018. Reducing language to rhythm: Amazonian Bora drummed language exploits speech rhythm for long-distance communication. *Open Science* 5(4). 170354. <https://doi.org/10.1098/rsos.170354>
- Seifart Frank, Doris Fagua, Jürg Gasché & Juan Alvaro Echeverri (eds). 2009. *A multimedia documentation of the languages of the people of the center. Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions*. Nijmegen, The Netherlands: The Language Archive. <https://hdl.handle.net/1839/00-0000-0000-001C-7D64-2@view>
- Woodbury, Anthony C. 2011. Language Documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.

Sonja Riesberg
sonja.riesberg@uni-koeln.de

Reflections on documentary corpora

Sally Rice
University of Alberta

For decades, language documentation proponents have argued for the separability of LD as its own sub-discipline. Many corpus linguists have made this same claim; thus, corpus linguistics shares the ethos of data over theorizing, whereby primary data represent authentic, connected discourse that is natural (not elicited), broadly sampled (across speakers, generations, dialects), and balanced (reflecting different usage contexts and genres). Nevertheless, many misconceptions remain about what a language corpus is, how it is formatted, how big or balanced it needs to be, and most importantly, *how it is queried*. In this reflection, I dispel some of these misconceptions, while reassuring community members and field linguists alike that a corpus is an exceedingly powerful tool for guiding the expansion of the documentary record, keeping precious language data in circulation, and helping to produce the classic descriptive by-products of LD such as dictionaries, phrasebooks, and grammars. Above all, the less-familiar but more direct by-products of corpus interrogation, such as word lists, frequency counts, concordance lines, N-grams, collocations, distribution, and dispersion plots, are so immediately interpretable and useful by speakers, learners, and linguists, that LD should give corpus linguistic training the same attention as project planning, ethics, recording, transcription, annotation, metadata, and archiving.

1. When documenting “linguistic practices” becomes focusing on actual spoken usage If the purpose of language documentation (LD), as so persuasively argued and succinctly crystallized by Nikolaus Himmelmann (1998: 166), is to provide “a comprehensive record of the linguistic practices characteristic of a speech community”, then a corpus is truly an excellent means of achieving this in ways readily accessible to speakers, learners, and outsider linguists. Indeed, Himmelmann wrote of the need to compile a collection or corpus of “communicative events”, recognizing, if only tacitly, that LD typically transpires in the context of orality; thus, spontaneous interactive conversation should be the centerpiece of documentary efforts. Himmelmann’s original articulation two decades ago (echoed and amplified by Woodbury 2003) of how documentary linguistics might especially focus on a different kind of primary data—

connected, naturally-occurring speech, whether narrative or conversation—dovetails more or less with recognition among corpus linguists that spoken language constitutes an equally important and thoroughly different mode of language use than that found in written genres. In the 1980s and 1990s, large national corpora for major languages like English pushed hard to include transcribed samples of spoken varieties alongside more easily compiled textual samples from newspapers, fiction, and academic writing. Insights about the profound differences between spoken and written modalities of language ensued (cf. the magnificent *Longman Grammar of Spoken and Written English*, Biber et al. 1999, based on the Longman Corpus Network corpora described at www.global.longmandictionaries.com/Longman/corpus); true corpus-based grammars and dictionaries of multiple languages also followed, as did new varieties of corpora, including learner, parallel, conversational, and multimodal corpora.

Despite the increasing recognition of the role that corpora play in LD and linguistics generally, there remain some entrenched misapprehensions about what a language corpus is and what one can do with such a corpus (be it big or small, balanced or skewed, annotated or not). In this reflection, I applaud the increase in calls for corpus-building in the LD and field linguistics literature (§2), spell out some of the prevailing misconceptions about language corpora in LD circles from the viewpoint of corpus linguistics proper (§3), and put the well documented “front-end” challenges of *building* a corpus in the first place (§4) alongside some of the many “back-end” benefits of *using* a corpus in the second place (§5). (Note, I intend *front-end/back-end* to be meant temporally, not in typical computational parlance of accessible/inaccessible to the user.) Chief among these benefits is getting a broader and sharper picture of actual spoken language usage patterns and patterns of variation within a speech community, a picture that can help inform subsequent stages of documentation.

2. Singing the virtues of documentary corpora: A rising chorus Since the publication of Himmelmann 1998, there has been a steady increase in edited volumes, textbooks, and handbooks about field linguistics and language documentation. Table 1 provides a list of some of the major book-length publications of the past two decades, arranged chronologically and showing the number of pages in the index under the heading *corpus/corpora* and the percent this represents against the total page number in each volume—an admittedly poor metric of attention, given the high degree of variability in indexing specificity and practice.

The notion of building documentary corpora is evidently growing more prevalent in the LD literature; see the steady upwards trend line in Figure 1, which graphically represents the percent frequency of mention of the words *corpus* or *corpora* by page in the volumes listed in Table 1. Sadly, it is still rare to find any listing for *conversation*, *speech*, or *interaction* in the typical LD index—the usual source of the primary data supposedly feeding into documentary corpora.

While it is heartening to see the role of corpora in LD being increasingly recognized (cf. McEnery & Ostler 2000; Scannell 2007; Mosel 2014), problematized (cf. Johnson 2004; Cox 2011; Jung & Himmelmann 2011; Vinogradov 2016), and evaluated (cf. Thieberger et al. 2015; Thieberger 2016), the field has a long way to go in understanding what a corpus is and is not. Moreover, the LD use of the word *corpus* as in *documentary corpus* is quite different from how a corpus linguist views the term. The focus in LD is generally on compiling the corpus, giving short shrift to what to do with the corpus data so compiled.

	Title	Corpus pages	Total pages	%
A	Newman & Ratliff (eds.) (2001). <i>LF</i> .	0	288	0%
B	Hinton & Hale (eds.) (2001). <i>Green Book of LR in Practice</i> .	1	468	0.2%
C	Gippert, Himmelmann, & Mosel (eds.) (2006). <i>Essentials of LD</i> .	47	424	11%
D	Crowley (2007). <i>FL: A Beginner's Guide</i> .	3	202	1.5%
E	Bowern (2008). <i>LF: A Practical Guide</i> .	8	285	3%
F	Grenoble & Furbee (eds.) (2010). <i>LD: Practice & Values</i> .	31	340	9%
G	Austin & Sallabank (eds.) (2011). <i>Cambridge Handbook of EL</i> .	38	567	7%
H	Chelliah & de Reuse (2011). <i>Handbook of Descriptive LF</i> .	21	492	4%
I	Haig et al. (eds.) (2011). <i>Documenting EL</i> .	1	344	0.2%
J	Thieberger (ed.) (2012). <i>Oxford Handbook of LF</i> .	38	545	7%
K	Sakel & Everett (2012). <i>LF: A Student Guide</i> .	2	179	1%
L	Jones & Ogilvie (eds.) (2013). <i>Keeping Languages Alive</i> .	6	269	2%
M	Jones (ed.) (2015). <i>EL & New Technologies</i> .	30	211	14%

Table 1: Number of pages in major LD publication indices mentioning corpus/corpora as a percentage of total pages overall. Volumes are listed chronologically. EL=endangered languages; FL=field linguistics; LD=language documentation; LF=linguistic fieldwork; LR=language revitalization.

3. Lingering misconceptions about what a corpus is Since Himmelmann 1998 first distinguished linguistic description and language documentation, the latter has become associated with collecting primary data in the form of audio and video recordings, making transcriptions and other annotations of such recordings, and compiling these transcribed representations, with appropriate metadata, into a corpus for archiving. Himmelmann's own view on using a documentary corpus suggests that it has "at least the potential of being of use to a larger group of interested parties. These include the speech community itself, which might be interested in a record of its linguistic practices and traditions" (ibid.: 163). This is an exceedingly vague and uninspiring illustration of the application of a documentary corpus. Indeed, the bulk of this seminal article is about corpus compilation, from the sampling of a full array of communicative event types to the metadata annotation that primary recordings and secondary transcriptions should receive.

Himmelmann 1998 definitely set the stage for codifying what I'm calling the *front-end* protocols of documentary linguistics: speaker sampling and recording techniques, transcription and annotation, metadata management and archiving. This much-needed attention has continued through Thieberger & Berez 2012 and just about every volume listed in Table 1. Unfortunately, these sources are usually replete with elusive and ultimately off-hand comments that do little to clarify exactly what a corpus is capable of. In several instances, the quotes in (1) exhaust the topic of *corpus* in their respective

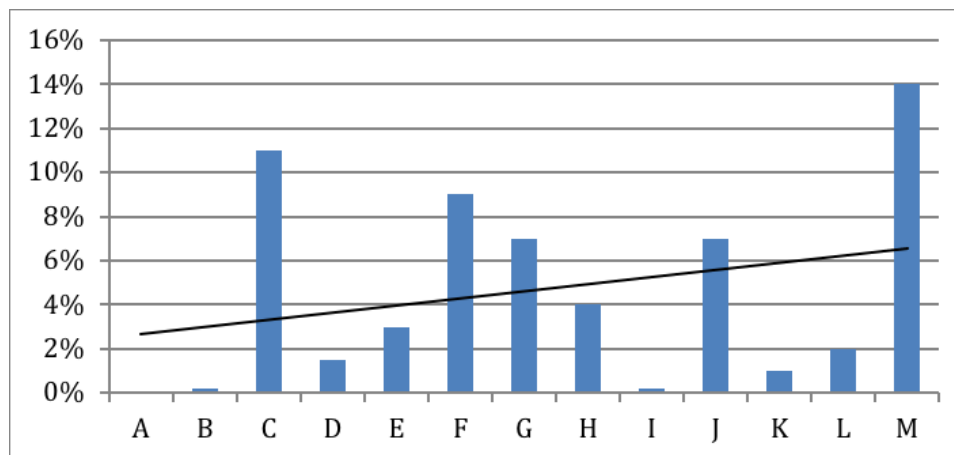


Figure 1: Slight but steady increase (see trend line) in percentage of pages in major LD publication indices mentioning corpus/corpora from 2001-2015, as listed chronologically and described in Table 1.

sources. Critically, they lead nowhere that an uninformed fieldworker never exposed to corpus linguistics can follow.

(1) Some minimalist comments about *using* a corpus in the LD literature

- a. “Corpus data is more useful if it’s annotated. That allows you to search for more detailed environments. It also allows you to create sub-corpora...that would let you search for differences between...two genres.” (Bowern 2008: 120);
- b. “Even given a large corpus of data, we may not have enough information to interpret the data without analysis. Thus, it is difficult to know whether all the linguistic forms and structures have been represented by the available data, whether paradigmatic gaps are intentional or rare, and what types of linguistic elicitations are needed to fill out the corpus of data.” (Berge 2010: 54);
- c. “The corpus should be annotated in a way that would allow a philologist in the distant future to interpret its content.” (Good 2010: 126);
- d. “There is absolutely no reason why the kinds of corpus-based statistical studies that have been carried out extensively on different varieties of English could not be carried out in other languages as well.” (Crowley 2007: 18);
- e. “Corpus linguistics does not typically result from the activities of fieldworkers, since corpora typically consist of written data easily studied by computational methods, although they are increasingly transcripts from spoken data.” (Chelliah & de Reuse 2011:12);
- f. “A well-formed corpus allows us to seek answers to linguistic questions that are difficult to ask when data is limited to what can be expressed on the printed page.” (Thieberger & Berez 2012: 116).

Indeed, in otherwise excellent overviews of creating and annotating language corpora, Vinogradov (2016) and Gries & Berez (2017) compare some basic characteristics of classic by-products of LD such as a Boasian text collection in terms of a variety of features, as shown in Table 2.

I have highlighted the last two features, *searchability* and *quantitative analysis*, because both are left as casually referenced and unexplained as the activities listed above in (1). Any corpus, however large or small, affords a birds' eye view of the material therein. It is this ability to search materials in the aggregate that allows the emergence of language-specific patterns that go beyond the anecdotal. Indeed, depending on the size of the corpus, some observations about pattern frequency can be statistically confirmed through simple association measures or openly challenged with more data. These patterns may be very fragmentary and low-level, but as recurrent expressions they generally constitute the core of actual language-in-use.

The heart of the matter is this: Suppose you're a middle-aged (or older) field linguist who came of age before the emergence of corpus linguistics or suppose you're an undergraduate or graduate student being trained in LD at a university that doesn't offer corpus linguistics training (which still describes the majority of linguistics departments)? How are you to square the circle between corpus creation and corpus application if you have never worked with a concordancer (the generic name for corpus-querying software), never queried multiple corpus files at the same time, never found strange patterns of co- or non-occurrence, never been surprised by the large number of fixed expressions that turn up, or never really confronted the staggering differences in frequency between lexical and grammatical material in a language or the idiosyncratic distribution of particular words or phrases in different genre types? Understanding what a corpus is and what it can do is only going to enhance and motivate the LD process itself. Going forward, we must stop regarding the corpus as a body of recordings, impeccably textualized and identified, and possibly left silent and still in an archive, but instead view it as an active and noisy collection of transcribed conversations teeming with insights about the language and its use that we can eavesdrop on again and again.

4. What a language corpus—documentary or otherwise—really is A corpus is neither a field linguistic database (as in a FLEx-style project with elicited fieldnotes, interlinearized utterances or narratives, a morpheme and word lexicon, etc.) nor a text collection. At its most basic, a corpus is a machine-readable collection of text *files* that can be queried simultaneously or selectively. In the case of spoken corpora—as documentary corpora are most likely to be—those digital text files will consist of transcriptions of speech (the output of transcription software such as ELAN, which allows for time-aligned annotation of an audio/video signal). If the speech source reflects unplanned conversation, then there will likely be incomplete utterances, repetitions, hesitations, interruptions, over-speech, all segmented into turns or intonation units. If the speech source reflects more planned narrative (a personal story, traditional legend, or oratory), then the transcribed text file may evidence more holistic, sentence-like structures. Regardless, both broad types of spoken language share the virtue of being natural and contextualized. Together with other communicative event types, they can form a corpus of mono- and dialogic language use as recorded in a speech community. Since the transcription (text) files are backed up by media as well as copious metadata, they themselves need not and should not contain any other information beyond the

Feature	Major corpora	Documentary corpora	Language archives	Printed text collection
selectivity of material	+	+	-	+/-
machine-readable format	+	+	+	-
volume	big/huge	small	big/small	very small
annotation	+	+/-	-/+	+/-
balanced subcorpora	+	-/+	-/+	-
searchability	+	+	-/+	-
quantitative analysis	+	-/+	-	-

Table 2: Presence or absence of basic characteristics of different research instruments for LD (adapted from Vinogradov 2016: 136 (his Table 3) and Gries & Berez 2017).

transcription and possibly an identifier for the speaker at each interactional turn. The text files that constitute the language corpus are not the same as the transcription files.

Here, I switch to a new moniker, language corpus or LC, to distinguish it from the documentary corpus, DC, that not only means something else, but is too often associated with inexact and promissory applications. A DC is about collecting and cataloguing data. An LC is about effectively and imaginatively exploring those data. Any time-stamped transcriptions, which may be further parsed, interlinearized, tagged, lemmatized, and translated into another language with attendant situational metadata, belong in the archived DC. Ultimately, the LC should be monolingual (code-switching aside). It can also be small, unbalanced, un-annotated and un-lemmatized (no reduction of inflected or derived forms to their bare stem). This lack of mark-up beyond a clean and consistent transcription is especially relevant in the earliest stages of LD when data are scarce, analytical knowledge is lacking, and the time and energy to annotate are in short supply (cf. Boerger 2011). Whereas these limitations can cripple language description and analysis, they constitute virtues in certain corpus linguistic camps, such as the neo-Firthians or the Birmingham School (cf. Sinclair 1991 and, especially, McEnery & Hardie 2012, Chapter 6, for helpful overviews), which regard corpus returns of un-annotated text or speech samples rather than linguistic theory or typological/areal expectation as the ultimate arbiter of what's going on in a language.

A real LC is not an archive of available material. It involves the rendering of that material to be machine-readable and query-able. In short, the LC is a folder, stratified or not, composed of a set of appropriately named text files. These files should have transparent file names that identify attributes deemed relevant to the particular LD project (e.g. speaker ID, genre, dialect, recording date, link to media file, etc.). Concordancers return data from queries linked to their source files, so good file-naming (the only place that metadata should reside) is especially pertinent during actual corpus searches.

5. What a language corpus can do (the neglected back-end) Thus far, I've lamented how applications of a corpus are left implicit in much of the LD literature. It's now time to be explicit and put a sample demonstration corpus through its paces. In (2) and (3), I list some common concordancer tools and corpus linguistic applications. The screen shots illustrated in Figures 2–7 are taken from Rice & Thunder 2017 and reflect data from a nearly 9,000-word corpus of *nêhiyawêwin* (or Plains Cree; ISO 639-3: cre), an Indigenous language of Western Canada, comprised of nine files representing three genres: casual conversation (C), planned narrative (N), and written stories (S). The demonstration concordancer into which the nine files were uploaded is AntConc (Anthony 2018), which can handle UTF-8 encoded (Unicode) plain text files and even help identify inconsistencies in spelling or file-rendering when files are first uploaded.

(2) Some classic concordancer tools

- a. orthographic or frequency-based word lists, as in Figure 2;
- b. keywords-in-context (KWIC), also known as concordance lines, as in Figure 3;
- c. N-grams or recurrent fixed expressions of various lengths, as in Figure 4;
- d. collocates of an item, be it morpheme, word, or expression, as in Figure 5;
- e. dispersion plots (which locate where in a file a certain string, be it morpheme, word, or phrase appears), as in Figure 6;

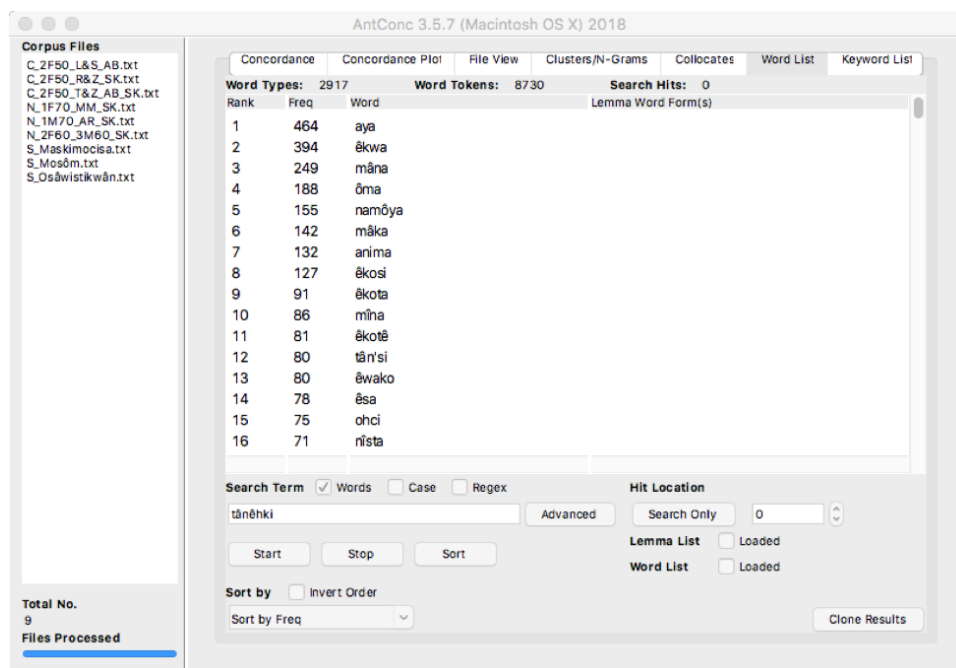


Figure 2: A frequency-based word list returned from the *nêhiyawêwin* demonstration corpus using the AntConc **Word List** function. Knowing which words are highly recurrent versus rare or absent in a corpus or in particular corpus files helps both the linguist and the language instructor target phenomena to investigate or teach. Here, the top-ranked word *aya* is a hesitation device. The next most frequent item *êkwa* ‘and’ is a conjunction.

- f. regular expression (regex) searches, using wildcards and other simple scripts to search within or across words, as in Figure 7.
- (3) Some classic corpus applications
- a. build exemplified dictionaries and grammars by providing a source of natural, example sentences (via concordance lines);
 - b. help with synonymy differentiation;
 - c. allow for sense disambiguation;
 - d. provide context for discoveries about semantic prosody;
 - e. demonstrate genre/register/dialect/gender/generational differences;
 - f. identify useful recurrent expressions, formulaic language, or phrasemes (as discussed in Rice 2017) that can help learners begin to develop conversational skills.

AntConc 3.5.7 (Macintosh OS X) 2018

Concordance Hits: 30

Hit	KWIC	File
1	itakwāw tǎn'si ehisiṁāmitoneyihtān	mistahi aniki māna nista nimāmitoneyir N_1M70_AR_SF
2	amākosiyān [laughter] ēkosi anima āw	mistahi anima nista kinanāskomitin kīsti C_2F50_R&Z_S
3	ēkotē ēkotē cī ēkwa cī ēkota	mistahi atoskēwin ehitaḱok tānitwa atoc C_2F50_T&Z_AI
4	yask kapiminākatohkātihcīk tāp'wē ēsa	mistahi atoskēwin kāwasoyan mihcētṽ C_2F50_R&Z_S
5	ltošisiminān ēkwa māna kāpēkiyokācīk	mistahi aya nikimiywēyihētēn ahpō ēp C_2F50_L&S_AI
6	ākpamohtahāyāhḱ ēkota aya ēkota aya	mistahi aya sōniyāw kihispayiw ēohpikil C_2F50_L&S_AI
7	wan nohtāwiy ēhatoskēt namōya māka	mistahi aya sōniyāwa aya ēhosihāt nikīr C_2F50_L&S_AI
8	siwikamīkohk oh miywāsin ēkosi māka	mistahi kinanāskomitin ētēpēyimoyan k C_2F50_R&Z_S
9	kiyām mistahi kāhitēyimisot ēwako ana	mistahi kāhitēyimisot mah pēyakwan m C_2F50_T&Z_AI
10	nisiṁis wiya ēwako ana kākikē kiyām	mistahi kāhitēyimisot ēwako ana mistaf C_2F50_T&Z_AI
11	ēhatāwēyāhḱ mičiwin ēkwa mīna māna	mistahi kākikē ēkōsēsāwihcīk aya pēyak C_2F50_L&S_AI
12	w onekihiḱomāwak ewako mīna iyīkohk	mistahi kā mākwīkoyahḱ aw ēkota kāpāi N_1M70_AR_SF
13	wihtamawāt kikway kīspin miywāsinīyiw	mistahi kikway ewanihtēyāhḱ ōma nehij N_1M70_AR_SF
14	isk māna kimiywēyihthen etikwe māna	mistahi kikway nimīciwin ekīhasahkecīk N_1F70_MM_SF
15	Yup	mistahi kikway āta mīna nista nimāmitoi N_1M70_AR_SF
16	in toni aya nipimikiskēyihētēn ēcika ōma	mistahi kikway ēpimikiskinoḱamākwiyī C_2F50_R&Z_S
17	emisiwanahtākamīkīsit peyakwan isko	mistahi māna kikway nimāmitoneyihthen N_1M70_AR_SF

Search Term: Words Case Regex

Search Window Size: 50

Buttons: Start, Stop, Sort, Show Every Nth Row: 1

Kwic Sort: Level 1 1R Level 2 2R Level 3 3R

Total No. 9
Files Processed

Figure 3: A set of concordance lines returned from a search of *mistahi* ‘a lot of’ sorted by first, second, and third word to the right using the Concordance function. From Hit lines 13-16, we can see that the phrase, *mistahi kikway* ‘a lot of something’, appears four times in the corpus across four distinct files: three narratives and one conversation. If one knew nothing about the language, the prevalence of this bigram in such a small corpus would suggest that it has some sort of unit status as an expression.

AntConc 3.5.7 (Macintosh OS X) 2018

Concordancer Concordance Pio File View Clusters/N-Gram Collocate Word Lis Keyword Lis

Total No. of N-Gram Types 48 Total No. of N-Gram Tokens 108

Rank	Freq	Range	N-gram
1	6	3	niya wiya aya
2	4	2	tânêhki mâka ôma
3	4	2	ékwa mâna aya
4	3	2	mâna aya tân'si
5	3	2	ékosi isi anima
6	3	2	ékwa aya aya
7	3	2	êwako anima aya
8	2	2	ah tân'si ôma
9	2	2	askaw mâna aya
10	2	2	askaw mâna piko
11	2	2	aya tân'si anima
12	2	2	aya tân'si nîsta
13	2	2	aya tân'si ôma
14	2	2	aya um namôya
15	2	2	kiya tân'tê ohci

Search Term Words Case Regex N-Grams Advanced

N-Gram Size Min. 3 Max. 3

Start Stop Sort

Sort by Invert Order Search Term Position Min. Freq. 2 Min. Range 2

Sort by Freq On Left On Right

Clone Results

Total No. 9 Files Processed

Figure 4: A set of 3-grams with a frequency of at least 2 and a range (number of files) of at least 2 returned using the **Clusters/N-Grams** function. This function can indeed launch a fishing expedition. We are asking the corpus to look for patterns of three recurrent words without any preconception as to their meaning or structure. In this case, of the 15 visible returns in the list, *aya* (a hesitation device), surfaces in 10 or $2/3^{rds}$ of the cases. In text or prepared narrative, any recurrent multi-word strings would likely be more informative and point to actual fixed expressions in the language.

AntConc 3.5.7 (Macintosh OS X) 2018

Concordanc Concordance Plr File Vie Clusters/N-Gram Collocate Word Li: Keyword Li

Total No. of Collocate Types: 15 Total No. of Collocate Tokens: 52

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	3	1	2	7.01139	so
2	2	2	0	5.91186	nimiywēyih̄tēn
3	2	1	1	4.32689	mistahi
4	2	2	0	4.23378	mmhhmm
5	5	2	3	4.21586	ēkotē
6	13	6	7	3.97422	māna
7	2	0	2	3.94838	yup
8	2	1	1	3.87623	isi
9	3	2	1	3.81875	wiya
10	2	0	2	3.80752	niya
11	4	2	2	3.24510	ēkosi
12	2	2	0	2.91186	ēwako
13	2	2	0	1.95766	namōya
14	5	3	2	1.93366	ēkwa
15	3	1	2	0.96077	aya

Search Term Words Case Regex Window Span Same
 kākikē Advanced From... 2L To... 2R
 Start Stop Sort
 Sort by Invert Order Min. Collocate Frequency 2
 Sort by Stat Clone Results

Total No. 9 Files Processed

Figure 5: Collocates of *kākikē* ‘always’ within 2 words to the left or right with a frequency of at least 2 and a range of at least 2 returned using the **Collocate** function. This function gives an indication of words that tend to co-occur within a fixed span, even though they might not be adjacent, such as *very* and *indeed* in varieties of English which frequently surface with an intervening adjective or adjectival phrase of varying length.

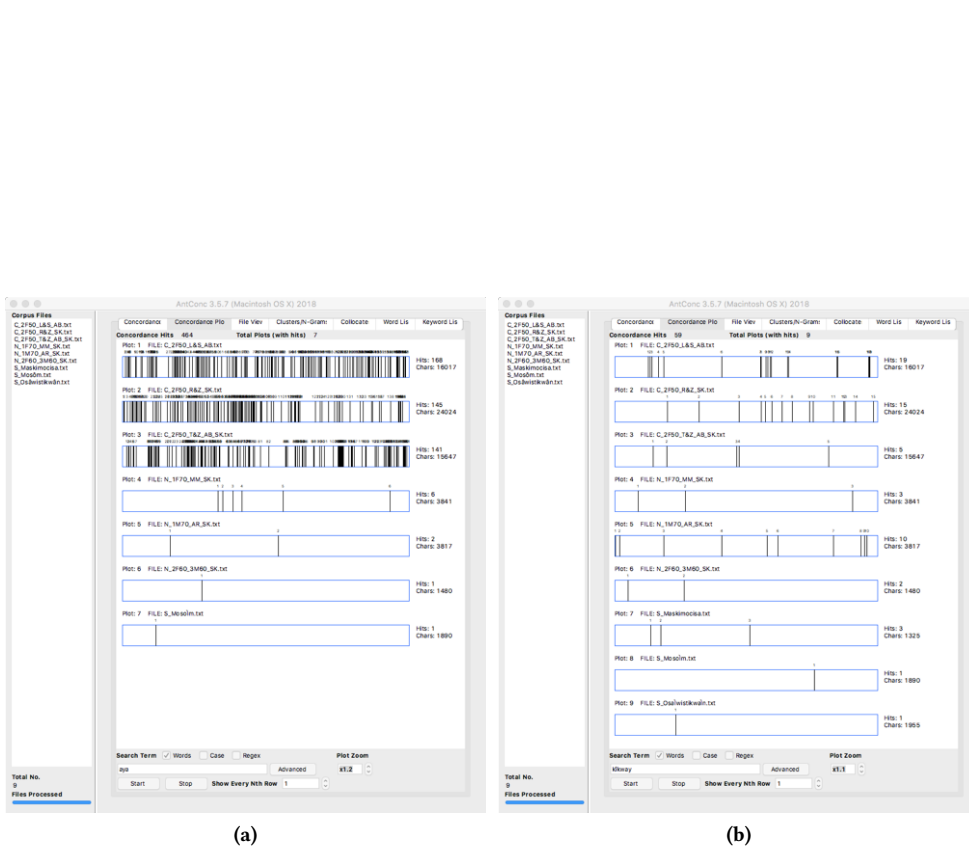


Figure 6: Two sets of dispersion plots for the hesitation device, *aya*, in (a) and the indefinite pronoun, *kikway*, in (b) showing, respectively, the highly skewed or relatively well distributed occurrence of each item within each corpus file. These results were returned using the **Concordance Plot** function. This function can yield immediate insights into differences in genre, speaker, etc., as well as differences in words that have more of a lexical vs. more of a grammatical function in the language.

AntConc 3.5.7 (Macintosh OS X) 2018

Corpus Files
 C_2F50_L&S_AB.txt
 C_2F50_R&Z_SK.txt
 C_2F50_T&Z_AB_SK.txt
 N_1F70_MM_SK.txt
 N_1M70_AR_SK.txt
 N_2F60_3M60_SK.txt
 S_Masimocisa.txt
 S_Mosom.txt
 S_Osawistikwân.txt

Concordance Hits 17

Hit	KWIC	File
1	pimitisahamân pâmayes mâcihatoskëyân ahkosiwikamikohk mihoët mihcêët nikpêht	C_2F50_R&Z_S
2	nâkwikoyahk aw êkota kâpâkisiniyak êkâ ekmâmawiwîchitoyahk êkâ wîchikoyâhk	N_1M70_AR_SF
3	cik apprehension order êwîmaskamihcik iyikohk mâna iyikohk êtâcîkwêcîk iskwe	C_2F50_R&Z_S
4	nitëyihêtên mm mmm âyiman cî kîspin kawanihtâyahk nêhiyawêwin yup toni tâp'	C_2F50_T&Z_AI
5	cohikoyahk niwâpahten nîsta ehitohteyân kiskinohamâtowikamikohk mâna konta mi	N_1M70_AR_SF
6	imâtisîwinaw êkîhsâkihtâk êkohchianîma kâkîkiskinohamâtahk êkwa êpakosëyimitz	C_2F50_R&Z_S
7	kôpitamêk kîhtwâm êhitiyit êsa ôma ôma kâpakwâtahk kayâs êsa mâna T kâhisit ayê	C_2F50_R&Z_S
8	â kiyânaw êkîpêhîsi paminikawiyahk kâpê kâpêhâpîsîsiyahk tânisi ôma kapimihîsînh	C_2F50_R&Z_S
9	nikîpe nikîpe ohcimikawînan kîkway ôma kâwîhkohtohk ôma epehohtpikiyâhk ehapi	N_2F60_3M60
10	ikwê ôma um 1985 kâwênîkîpêkîwân ôta maskwacîsihk êkwa êkotê nikîmâthatoski	C_2F50_L&S_AI
11	i Lena Lapatack nitîsîyîhkâson niya êkwa onihcikîswapiwînihk ohci êkwa kiya tân'tê	C_2F50_L&S_AI
12	îhakîk nân'taw kêtôhtêcîk namôya kâkîkê pihcâyîhk kâyêcîk askaw mâna aya mihcê	C_2F50_T&Z_AI
13	îtawîhayamîhthîkîcîk kîspin kitamîsihîcîk schoolîhk nân'thaw kesimamayîhthîkîcîk	N_1F70_MM_SI
14	Wâpahtamwak nama kîkway e-asîwateyîk wîyâkanîhk êkwa e-pîkopayîniyîk tehtapîk	S_Osawistikwâr
15	wê kiyânaw kakiskinohamawayakîk tânisi êhisiwahkohtoyahk we got carried away a	C_2F50_L&S_AI
16	êwin pihtwâwin êkosi isî ohipimê nânîkaw êkihçëyîhtamîhk kakanawëyîmiyît mâka k	C_2F50_R&Z_S
17	kâmâmîskôhtamân êkîmâmawîpayiyâhk êkîmâmawîtohtamahk kîkwaya ahpô ôma	C_2F50_L&S_AI

Search Term Words Case Regex Search Window Size 50

Show Every Nth Row 15

Kwic Sort Level 1 0 Level 2 0 Level 3 0

Total No. 9
Files Processed

Figure 7: Concordance lines returned from a regex search using the **Concordance** function. Although there are three allomorphs in *nêhiyawêwin* of the locative suffix, {-*ihk*, -*ohk*, -*ahk*}, words ending in all three variants can be queried simultaneously using a regular expression such as `\w+(i|o|a)hk\s`. The use of regular expressions when conducting corpus searches helps overcome challenges caused by allomorphy, variation in spelling, incomplete knowledge, or other context effects that may affect a form.

Two widely subscribed LD maxims are also shared by corpus linguists: (i) taking a language as it comes (not based on translation, elicitation, or someone else's analysis) and (ii) making samples of language accessible and re-useable for multiple purposes and users. If we must all do as much as we can with the language samples we've got, then the multiple queries that can be conducted on a language corpus by a concordancer seem downright economical and efficient. There is huge bang for the corpus buck, in both early and late stages of LD. Seeing data displayed in the form of corpus returns also serves as an inspiration and a directive to collect more samples more broadly from more usage situations and speakers, if at all possible. A small, untagged, and unbalanced corpus can still yield tremendous insights into the structure, meaning, and use of a language—sampling skews never go away, regardless of corpus size. Most endangered language communities or LD projects led by a single individual probably have all the tools and personnel needed to start building and using a language corpus. The creation and maintenance of such a corpus can involve a variety of community members with differing skills and interests, from recording and transcription to file-editing and metadata management (cf. Boerger 2011). Community-led, corpus-based LD projects can go hand-in-hand without much or any intervention from a linguist or programmer. Amongst the many new skill sets that field linguists and endangered language activists need to develop—beyond linguistic analysis, ethical conduct, grant-writing, and front-end protocols—should be a basic understanding of corpus linguistics.

In re-conceptualizing the documentary corpus as an actualized, query-able corpus of everyday conversation or communicative events, the benefits of corpus-creation and the bounties afforded by interrogating such a corpus with proper concordancing tools can be explicitly demonstrated, demystified, and hopefully implemented widely by speakers and learners in endangered and minority language speech communities. With the availability of free, off-the-shelf, easy-to-use, Unicode-savvy, XML-capable, multi-platform, 4th generation (stand-alone) concordancers such as AntConc, a corpus does not have to live on-line, but can reside on a computer (or two) in a community. Thus, LD and documentary corpora will be able to achieve a few of the widely held desiderata itemized by Bird & Simons 2003, Woodbury 2003, Himmelmann 2006, and others: a lasting, multipurpose, and re-useable product of documentary efforts. The LD field has spent enough time talking about coverage. It's time to leverage that coverage into actually applying corpus tools and conducting corpus analyses that allow precious language data to speak for themselves without descriptive or analytic overlay and, most importantly, without further delay.


References

- Anthony, Laurence. 2018. AntConc: A freeware concordance program for Windows, Macintosh OSX, and Linux (Version 3.5.6) [Computer Software]. Tokyo: Waseda University. Available from <http://www.laurenceanthony.net/>.
- Austin, Peter K. & Julia Sallabank (eds.). 2011. *The Cambridge handbook of endangered languages*. Cambridge: Cambridge University Press.
- Berge, Anna. 2010. Adequacy in documentation. In Grenoble & Furbee (eds.), *Language documentation: Practice and values*, 51–66. Amsterdam/New York: John Benjamins.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.
- Bird, Steven & Simons, Gary. 2003. Seven dimensions of portability for language documentation. *Language* 79(3). 557–582.
- Boerger, Brenda. 2011. BOLDly go where no one has gone before. *Language Documentation & Conservation* 5. 208–233.
- Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*, 1st edn. London/NY: Palgrave MacMillan.
- Chelliah, Shobhana & Willem de Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. Dordrecht: Springer.
- Cox, Christopher. 2011. Corpus linguistics and language documentation: Challenges for collaboration. In John Newman, R. Harald Baayen & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 239–264. Amsterdam: Brill.
- Crowley, Terry. 2007. *Field linguistics: A beginner's guide*. Oxford: Oxford University Press.
- Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.
- Good, Jeff. 2010. Valuing technology: Finding the linguist's place in a new technological universe. In Grenoble & Furbee (eds.), *Language documentation: Practice and values*, 111–131. Amsterdam/New York: John Benjamins.
- Grenoble, Lenore A., N. Louanna Furbee (eds.). 2010. *Language documentation: Practice and values*. Amsterdam/New York: John Benjamins.
- Gries, Stefan Th. & Andrea Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 379–409. Dordrecht: Springer.
- Haig, Geoffrey, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.). 2011. *Documenting endangered languages: Achievements and perspectives*. Berlin: de Gruyter.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.
- Hinton, Leanne & Hale, Ken (eds.). 2001. *The green book of language revitalization in practice*. New York: Academic Press.
- Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.), *Language documentation and description*, vol. 2, 140–153. London: SOAS.

- Jones, Mari C. (ed.). 2015. *Endangered languages and new technologies*. Cambridge: Cambridge University Press.
- Jones, Mari C. & Sarah Ogilvie (eds.). 2013. *Keeping languages alive: Documentation, pedagogy, and revitalization*. Cambridge: Cambridge University Press.
- Jung, Dagmar & Nikolaus P. Himmelmann. 2011. Retelling data: Working on transcription. Haig, Geoffrey, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.), *Documenting endangered languages: Achievements and perspectives*, 201–220. Berlin: de Gruyter.
- McEnergy, Tony & Andrew Hardie. 2012. *Corpus linguistics: Methods, theory, and practice*. Cambridge: Cambridge University Press.
- McEnergy, Tony & Nick Ostler. 2000. A new agenda for corpus linguistics – working with all of the world’s languages. *Literary and Linguistic Computing* 15(4). 403–420.
- Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. *Language Documentation & Conservation* 8. 135–157.
- Newman, Paul & Martha Ratliff (eds.). 2001. *Linguistic fieldwork*. Cambridge: Cambridge University Press.
- Rice, Sally. 2017. Phraseology and polysynthesis. In Michael Fortescue, Marianne Mithun & Nicholas Evans (eds.), *The Oxford handbook of polysynthesis*, 203–214. Oxford: Oxford University Press.
- Rice, Sally & Dorothy Thunder. 2017. Community-based corpus-building: Three case studies. Paper presented at the 3rd International Conference on Language Documentation and Conservation. Honolulu, March 2-5, 2017.
- Sakel, Jeanette & Daniel L. Everett. 2012. *Linguistic fieldwork: A student guide*. Cambridge: Cambridge University Press.
- Scannell, Kevin P. 2007. The Crúbadán project: Corpus building for under-resourced languages. *Cahiers du Central* 5. 5–15.
- Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Thieberger, Nicholas. (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press.
- Thieberger, Nicholas. 2016. Documentary linguistics: Methodological challenges and innovatory responses. *Applied Linguistics* 37 (1). 1–13.
- Thieberger, Nicholas & Andrea Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.
- Thieberger, Nicholas, Anna Margetts, Stephan Morey & Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21.
- Vinogradov, Igor. 2016. Linguistic corpora of understudied languages: Do they make sense? *Kāñina* 40(1). 127–141.
- Woodbury, Anthony C. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description*, vol. 1, 35–51. London: SOAS.

Sally Rice

srice@ualberta.ca

 orcid.org/0000-0002-2988-321X

Reflections on the role of language documentations in linguistic research

Stefan Schnell

University of Bamberg

Centre of Excellence for the Dynamics of Language

I reflect the role of language documentations in linguistic research beyond its most common linguistic use as a high-quality database for descriptive work. I show that the original Himmelmann-ian conception of documentations, as multi-varied and multi-purpose, and to some extent community-driven, enable a range of research outcomes that would not have been foreseeable within the traditional descriptive, typological and theoretical agendas. I argue that it is overall more fruitful for innovative linguistic research to invest into the processing of haphazard language documentation data rather than attempting to collect precisely the kind of data demanded by specific analytic goals.

1. Introduction¹ According to Himmelmann (1998) language documentations are ideally prepared not in service of any specific analytic agenda, but as broad multi-varied collections open to a variety of purposes and uses by different user groups, including speech communities and academic linguists. In this contribution I show in what ways data that are not controlled for any research purpose can play an important role in research outside traditional descriptive grammar writing. I show that documentation-based research always involves considerable efforts in additional or alternative processing of different types of data (Himmelmann 2012), but that it often does not need to involve further collection of more *specifically useful* data in order to play its increasingly important role in linguistics.

In what follows, I outline first how language documentations are seen to be empirically valuable (Section 2). I then summarize some prominent research agendas that traditionally

¹The ideas developed in this contribution have been developed while I was a postdoc researcher in the ARC Centre of Excellence for the Dynamics of Language. I am grateful for the ongoing support by the Centre. I would also like to thank Rebecca Defina, Cris Emmonds-Wathern, Jenifer Green, Yukinori Kimoto, as well as Lauren Gawne, Bradley McDonnell and one anonymous reviewer for helpful comments on an earlier draft of this paper. All remaining errors and shortcomings are my own responsibility.

build on rich performance data (Section 3). In Section 4 I turn to more recent research based on entire language documentations that has resulted in significant insights into language use across languages. In Section 5 I discuss the role of different text varieties in typologically oriented documentation-based research. I conclude my contribution in Section 6 with some reflections on further developments in corpus-based typology.

2. Empirical value of language documentations The most obvious and central value of language documentations for academic linguistics is that it provides an empirical basis on which linguistic analyses are accountable by way of giving access to the recorded data and their annotation (Himmelman 2006; also Gawne & Berez-Kroeker 2018 in this volume; Berez-Kroeker et al. 2018 for recent discussion). Aside from this more global value of accountability, different components of a documentation, "...a comprehensive record of the linguistic practices of a given speech community" (Himmelman 1998:166), are more or less useful in reflecting different aspects of these practices, in particular observable behavior and metalinguistic knowledge. Hence, Himmelman (1998) advocates aiming at a broad collection of data resembling different degrees of naturalness and spontaneity. Thus, a casual conversation recorded with little awareness of participants will be the best representation of naturally, and most frequently, occurring observable linguistic behavior in a speech community. Metalinguistic knowledge, on the other hand, is often not reflected in such recordings and specialized elicitations of, for instance, morphological paradigms together with comments on similarities of forms, etc would capture this instead, but not represent any naturally occurring speech event. Elicitation sessions, as well as other data that do not resemble any established communicative routines, like stimuli-based elicited texts, are on the other hand characterized by a high degree of spontaneity, which may reveal interesting aspects of a language systems otherwise not represented in more naturalistic data. A further dimension concerns the relevance of documentary activities for a given speech community: elicitations of a morphological paradigm are obviously not of any major concern for communities, but neither are more natural casual conversations. Instead, collection of different forms of verbal art, indigenous oral literature as well as encyclopedic knowledge of flora, fauna, material culture and so forth, and respective vocabulary, is often among the major desiderata of a given community (Himmelman 2006; Mosel 2014a).

Similarly, different types of data play different roles in linguistic research. Casual conversations resemble most accurately how a language is used at a given point in time in a community, and this data is ultimately crucial for a thorough understanding of language change and possible developments of evolutionary models thereof (e.g. Baxter & Croft 2016; Blythe & Croft 2012). It would be much less useful for a first descriptive account of a language which requires examples of complete, well-formed constructions which can be hard to come by in conversational data. Narrative texts from oral literature may be a much better data source for this purpose. Elicitations of specific structures can provide data most relevant for descriptions in the most immediate way, including data often not attested in any less controlled data type (see Evans 2008; Rhodes et al. 2006), but they are hardly ever really useful for studies in language variation and diversification. To what extent different types of data are restricted to very specific linguistic purposes or open to a variety thereof is discussed in detail in McDonnell (2018, in this volume). My main concern here is to show how documentations that are mainly concerned with the coverage of linguistic practices and the desires of speech communities can play and have played an important role in linguistic research beyond the traditional descriptive paradigm.

3. Traditional research on language use: in search for the right data Performance data have long been a focus of dedicated research traditions in linguistics, for instance variationist sociolinguistics (Meyerhoff 2010), general corpus linguistics (Biber & Conrad 2009), or conversation analysis (Seedhouse 2013), among many others. Of particular relevance for typological linguistics has been a line of research that DuBois (2017) calls “discourse and grammar”: established by Wallace Chafe and Talmy Givón in the 1970ies, it is concerned with patterns of reference and information packaging in discourse and seeks to explain these with reference to cognitive factors of language processing (e.g. Givón 1976; Chafe 1976; DuBois 1987 among many others). On the other hand, identified patterns in discourse are considered the seedbed of grammatical structures which *emerge* through frequent deployment of discourse patterns during communication, hence the emergentist credo that “grammars do best what speakers do most” (Du Bois 1985). Grammar and discourse is closely related to the tradition of *language variation and change*, where language-internal and -external (i.e. social, cultural, etc) factors are related to regularities of language use and resulting diachronic developments (Labov 1994; Croft 2000).

A major challenge for these research traditions has been to determine what kind of performance data is required, in line with their respective goals, for instance sociolinguistic interviews in sociolinguistics, etc. Finding appropriate performance data has been, and continues to be, a particular challenge in more typologically oriented research, like that in grammar and discourse: these research agendas require records of connected discourse from as many languages as possible, comparable, at least to a certain degree. A common response to this challenge is to use stimuli-based elicited narrative texts, most notably Chafe’s (1980) Pear Film, or Mayer’s (1969) Frog stories (see Slobin 2004), which ensure a minimal degree of comparability of different texts on each occasion of their elicitation. While such elicited texts can yield interesting observations on possible structures of a language system due to their high degree of spontaneity, they do not capture natural routines of linguistic performance (see Foley 2003 for critical discussion of the use of Frog stories in Watam). In extreme cases, they hardly resemble any kind of coherent discourse at all, as reported by DuBois (1980) for the elicitation of Pear stories in Sakapultek. It seems to me that for typologically oriented studies of language use, we are still exploring what the ideal dataset looks like, and in the following I will show that language documentations have a great deal to contribute to this quest.

4. A found treasure: language documentations in usage-based linguistic research

Although Himmelmann (1998) mentions potential uses of language documentations outside the standard grammaticographic line, specific research projects of this kind drawing extensively on documentation data started to take off not before about ten years ago or so. It is worth mentioning that the DoBeS program dedicated its final fully-fledged round of funding almost entirely to projects utilizing existing collections in broader research projects. Examples are Frank Seifart’s project² on the ratio of nouns, pronouns and verbs in spoken-language discourse, Anna Margetts’ project³ on three-participant

²DoBeS research project *The relative frequencies of nouns, pronouns, and verbs cross-linguistically* (PI Frank Seifart, 2012-2015),

³DoBeS research project *Cross-linguistic patterns in the encoding of three-participant events* (PI Anna Margetts, 2012-2016) and *Cross-linguistic patterns in the encoding of three-participant events—investigating BRING and TAKE* (PI Anna Margetts, 2017-2018)

constructions across languages, and Claudia Wegener's project⁴ on prosodic patterns in discourse structure in contact situations between languages from two different families. Not specifically funded by DoBeS, but developed in its context is Geoffrey Haig's and my own project on referential choice and argument realization in discourse (Haig & Schnell 2014, 2016a,b).

The great potential of documentation-based research lies in its focus on aspects of language production that are not typically part of structuralist descriptive and typological work and in the embeddedness of performance data in the cultural context of speech communities. Both of these aspects have to teach us a lot about how languages are used, and how this may influence their evolution. For instance, Himmelmann (2014) draws upon extensive spoken language data to bear on the long-standing challenge of explaining the suffixing preference in the languages of the world. His study identifies a systematic distribution of dysfluencies and pausing in spoken discourse that corroborates specifically constrained structural contexts for the development of affixal exponents of grammatical categories, hence explaining the typological preference. Seifart et al. (2018) show that across languages, the production of noun phrases affords more planning effort, the latter being determined by proxy measurements of pause probabilities and speech rate. The authors attribute this higher effort to the particular referential choices associated with noun phrase production, a conclusion of major relevance to questions of referential choice and language processing in general. Further prominently published documentation-based studies are Margetts (2015) and Haig & Schnell (2016b).

These examples bear witness of the fact that the role of documentation-based research is gaining ground in academic linguistics. They also seem to yield some methodological insights that are important for future developments: for one thing, documentation-based research of this kind involves considerable efforts of further processing of existing data. Seifart et al. (2018), for instance, draw on data with word-level time-aligned transcriptions and further annotations that required the development of forced (time-)alignment methods (Strunk et al. 2014). Haig & Schnell's (2016b) extensive corpus study on argument realization draws to a large extent on a multilingual corpus annotated for specific morphosyntactic and semantic features of syntactic arguments (Haig & Schnell 2014). This required the development and monitored implementation of annotation guidelines that are applicable to diverse languages. Similar kinds of annotation guidelines have been implemented in Margett's three-participant project (Margetts et al. 2017). These observations counteract occasional ideas that linguistic analysis could in some way just fall out of documentations, as long as these are well-structured and well-curated. Instead, analytical documentation-based work seems to always come with research-specific additional efforts of data processing.

For another thing, it seems obviously worth pursuing research on data that has not been collected for specific research goals, and has thus not been controlled for in relevant ways. The relevant aspect is that the data resemble real, usually spoken, communication between speakers of diverse communities, and this can be of any kind in order to yield relevant research findings.

5. The usefulness and utilization of original and introduced text varieties From a more corpus-linguistic perspective though, paying attention to the characteristics of

⁴DoBeS research project *Discourse and prosody across language family boundaries: two corpus-based case studies on contact-induced syntactic and prosodic convergence in the encoding of information structure* (PI Claudia Wegener, 2011–2013)

specific text varieties is vital. I will first discuss the example of a variationist study in Vera'a that draws on different types of documentation data, and then turn to recent developments in the field of corpus-based typology, where the use of stimuli-based versus original text data is a major concern.

5.1 Utilization of minimally varied corpora in variationist studies In a study of object realization in Vera'a (Oceanic, North Vanuatu) (Barth & Schnell 2018), we drew on a sub-corpus of the overall language documentation which resulted from a fairly typical documentation project within the DoBeS program.⁵ The alternation we were interested in was that between a pronoun versus zero as a form of realization for those objects that are not a full noun phrase. We investigated spoken narratives as well as descriptions of both floral and faunal species, so that the texts in our corpus resemble two different registers (narration, description) with three different ontological classes of global discourse topic (humans, fish, plants). We find that the best predictor for the use of a pronoun is the global discourse topicality of the referent in question, being either the human protagonists in stories, or the fish and plant species under discussion, thus refining Schnell's (2012) treatment of the alternation as an animacy effect. Only a cross-register analysis of this kind, together with the implementation of sophisticated statistical methodology, made it possible to disentangle the notoriously converging dimensions of animacy and global topicality. Again, this study involved a considerable amount of meticulous corpus annotation work with GRAID (Haig & Schnell 2011) and subsequent further coding of data. This annotated Vera'a corpus is being archived as part of Multi-CAST (Haig & Schnell 2014) with the *Language Archive Cologne* (LAC), thus ensuring reproducibility of this study in the sense of Gawne & Berez-Kroeker (2018, in this volume).

Our inclusion of descriptive texts in our corpus investigation was motivated by my fairly random observations of pronoun use during data processing. However, to collect such data in the first instance was not motivated by our study at all, but followed from the design of the preceding documentation project where a team of researchers from various disciplines and local language workers aimed to document a large range of communicative events and various cultural aspects of two speech communities, including encyclopedic knowledge and associated vocabulary (and folk taxonomies) of flora and fauna, material culture, social organization etc. In accordance with the interests of all participants, we collected not only oral literature (and produced written editions thereof, Vorës & Schnell 2012), but also descriptions of flora and fauna, and their names and taxonomy information. These collections served as a basis for dedicated community materials, akin to the materials for the Teop language of North Bougainville produced by Ulrike Mosel and collaborators (e.g. Mosel et al. 2010; Mosel 2014b,c). It is important to note that descriptive (as well as procedural) texts are not an established genre in Vera'a linguistic culture, and are in this sense not natural; relevant information is traditionally conveyed only by means of demonstration. Their collection was motivated entirely by the aim of documenting the ethnobiological knowledge contained therein. In conclusion, it is possible to arrive at typologically highly significant results by exploring those kinds of data that come up for different reasons during a documentation project.

⁵DoBeS documentation project *Documentation of Vurës and Vera'a, the two surviving endangered languages of Vanua Lava, Vanuatu* (PI Catriona Hyslop) and *Documenting biocultural diversity in the languages of Vurës and Vera'a* (PI Catriona Malau, 2009-2011).

5.2 Original and introduced text varieties in corpus-based typology The typological field most clearly concerned with language use is *corpus-based typology*, a relatively recently emerging field that seeks to determine cross-linguistic commonalities in language use as well as respective diversity. Some of the pioneering work in corpus-based typology continues (and considerably improves on) the Chafe-Givón tradition introduced above. For instance, Bickel (2003) and Stoll & Bickel (2009) take up the long-standing question as to whether speakers of diverse languages overtly realize all syntactic arguments (as in English), or tend to leave them zero (as in Japanese). Rather than considering specific grammatical rules that may constrain the occurrence of zero arguments, they employ corpus measurements that they call *referential density* or *lexical referential density*, respectively. Obviously, in order to arrive at a useful comparison of speakers' argument realization behavior one needs to compare texts of roughly the same content, since content will be a major factor determining whether a particular referent is familiar at a given point in discourse or not. To achieve this goal, Pear stories are used since here the stimulus ensures that different speakers, including those of different languages, will recount roughly the same content, having the same number of opportunities to verbalize specific referents.

Similar considerations motivate the use of the so-called *Family Problem Task* (San Roque et al. 2012) in the *Social Cognition* project (Barth & Evans 2017) which seeks to determine cross-linguistic differences in the realization of certain communicative tasks, for instance the expression of thoughts of others or reference to human beings. It enables comparison of relevant lexical and constructional choices by different speakers from different languages in precisely the same contexts, as determined by the structure of the stimulus. To compare such choices across speakers, languages, and different types of context will obviously not bear any useful insights.

A line of research where the use of stimuli-based data has proven to be problematic though is that of DuBois' (1987) famous hypothesis of *preferred argument structure* (PAS): based on a small corpus of Pear stories from the Mayan language Sakapultek (Guatemala), PAS has until recently been widely accepted as a usage-based account for ergative grammar in the world's languages, see Evans & Levinson (2009). Adducing a range of corpus data from different languages, Haig & Schnell (2016b) demonstrate that PAS does not seem to extend beyond this single corpus from Sakapultek, whose containing texts seem to be characterized by what Haig & Schnell (2016b) call a "telegraphic style", presumably due to the immense discomfort speakers experienced during the respective experiment, as reported by DuBois (1980). Moreover, Schnell (under revision) finds that patterns of referent introduction are much better explained by reference to the way characters are presented in the movie stimulus rather than universal cognitive constraints on information flow. Hence, the use of a single stimulus may bear analytical risks. The latter two studies draw to a large extent on corpus data from language documentations that are not controlled for content, but instead have the advantage of resembling much more closely the kinds of routines in language use and that are variable to some degree, so that respective findings are not entirely dependent on a single type of text data.

What kinds of discourse data should underlie corpus-based typological studies is thus not determined by general methodological principles but by specific requirements related to research design and goals. An undebatable general requirement that modern developments in corpus-based typology have made considerable progress in is the accountability of findings: Haig & Schnell (2016b) draw largely on an archived and web-accessible multilingual corpus, called Multi-CAST (Haig & Schnell 2016a), that enables scrutiny and replicability of their findings by other researchers. Likewise, the Social

Cognition corpus is going to be accessible via PARADISEC. This is a great improvement of earlier work where for instance Pear story corpora have almost never been made available. This is of course not to deny that the accessibility of language documentation as well could be improved considerably, see Gawne & Berez-Kroeker (2018, in this volume).

Haig & Schnell's (2016a) MultiCAST initiative as well as the Social Cognition project (and likewise Margett's 3-participant project) again involve tremendous efforts of project-specific data processing, adding further layers of specialized annotations triggering certain constructional variants and some semantic features. In all cases, the annotations are comparable in nature, being applicable to diverse languages and enabling cross-linguistic comparison of corpus analyses. These annotations are clearly an improvement over traditional variationist procedures where relevant information is typically added in separate spreadsheets. Combining various layers of data annotation, all time-aligned to the recorded signal, opens up unprecedented possibilities for further studies, as has partly been done where GRAID annotations (Haig & Schnell 2014) have been combined with other annotation, like Schiborr et al.'s (2018) *referent indexes* (Schnell et al. 2018).


6. Conclusions I hope to have shown here that documentation-based linguistic research has enormous potential to yield insights into language use and language systems that would not have been foreseeable from the perspective of established descriptive, typological or theoretical traditions. In this connection, the haphazard and often not academically driven nature of documentations can often be an advantage, since it may bring up data that would not have been planned for from an academic research point of view, but that nonetheless provides the most relevant insights, as in the case of Vera'a plant and fish descriptions. Neither would this data have been collected if researchers had followed a purist ideal of naturalistic data. Although some research questions clearly demand the collection of very specific data (for instance directly comparable text data), it seems to me that time and effort is probably better invested into further data processing of what is there rather than collection of data in service of specific analytical goals. In this way, language documentations have an important role to play the scientific research into human language.

References

- Barth, Danielle & Nicholas Evans. 2017. SCOPIC Design and Overview. In Danielle Barth & Nicholas Evans (eds), *The Social Cognition Parallax Interview Corpus (SCOPIC): A Cross -linguistic Resource* (Language Documentation & Conservation Special Publication 12), 1–23.
- Baxter, Gareth & William Croft. 2016. Modeling language change across the lifespan: individual trajectories in community change. *Language Variation and Change* 28. 129–173.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David L. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meyer, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. <https://doi.org/10.1515/ling-2017-0032>
- Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.
- Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.
- Blythe, Richard A. & Willam Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88(2), 269–304.
- Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 25–55. New York: Academic Press.
- Chafe, Wallace (ed.). 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: ALEX Publishing Company.
- Croft, William. 2000. *Explaining language change*. Harlow: Pearson Education.
- DuBois, John W. 1980. Introduction – The search for a cultural niche: Showing the Pear Film in a Mayan community. In Wallace Chafe (ed.). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*, 1–8. Norwood, NJ: ALEX Publishing Company.
- DuBois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax*, 343–366. Amsterdam & Philadelphia: John Benjamins.
- DuBois, John W. 1987. The discourse basis of ergativity. *Language* 63(4). 805–855.
- DuBois, John W. 2006. The Pear story in Sakapultek Maya: A case study of information flow and Preferred Argument Structure. In Mercedes Sedano, Adriana Boliva & Martha Shiro (eds.), *Haciendo Lingüística: Homenaje a Paola Bentivoglio*, 191–221. Caracas: Universidad Central de Venezuela.
- DuBois, John W. 2017. Ergativity in discourse and grammar. In Jessica Coon, Diane Massam & Lisa D. Travis (eds.), *The Oxford handbook of ergativity*, 23–58. Oxford: Oxford University Press.
- Evans, Nicholas. 2008. Review of Essentials of language documentation. *Language Documentation and Conservation* 2(2). 340–350.
- Evans, Nicholas & Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5). 429–448.

- Gawne, Lauren & Andrea Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation on the 20 year anniversary of Himmelmann 1998*, 22–32. (Special issue of *Language Documentation & Conservation* 15.) Honolulu: University of Hawaii Press. <http://hdl.handle.net/10125/24804>
- Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In Charles N. Li (ed.), *Subject and topic*, 149–188. New York: Academic Press.
- Haig, Geoffrey & Stefan Schnell. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Manual version 7.0. ms. (Available at: https://www.academia.edu/10328418/Haig_Geoff_and_Stefan_Schnell._2014._Annotations_using_GRAID_Grammatical_Relations_and_Animacy_in_Discourse_.Manual_Version_7.0)
- Haig, Geoffrey & Stefan Schnell. 2016a. *Multi-CAST: multilingual corpus of spoken annotated texts*. Cologne: Language Archive Cologne. (<https://lac2.uni-koeln.de/de/multicast/>)
- Haig, Geoffrey & Stefan Schnell. 2016b. The discourse basis of ergativity revisited. *Language* 92(3). 591–618.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.
- Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation* 6, 187–207.
- Himmelmann, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words. *Language* 90.4, 927–960.
- Labov, William. 1994. *Principles of linguistic change*. Oxford: Blackwell.
- Li, Charles N. (ed.). 1976. *Subject and topic*. New York: Academic Press.
- Mayer, Mercer. 1969. *Frog where are you?* New York: Dial Books.
- Margetts, Anna, Katharina Haude, Nikolaus P. Himmelmann, Dagmar Jung, Sonja Riesberg, Stefan Schnell, Claudia Wegener, John Hajek, Andrew Margetts. 2017. Investigating three-participant events across text corpora. Paper presented at the workshop *Advances in corpus-based typology: exploring corpora of semi-parallel and indigenous texts*, convened by Geoffrey Haig, Stefan Schnell, Nicholas Evans, Danielle Barth, at the *ALT12*, Canberra, A.N.U., 15 December 2017.
- McDonnell, Bradley. 2018. Reflections on linguistic analysis. In Bradley McDonnell, Andrea Berez-Kroeker, and Gary Holton (eds.), *Reflections on language documentation on the 20 year anniversary of Himmelmann 1998*, 191–200. (Special issue of *Language Documentation & Conservation* 15.) <http://hdl.handle.net/10125/24820>
- Meyerhoff, Miriam 2010. *Introducing sociolinguistics*, 2nd edn. London/New York: Routledge.
- Mosel, Ulrike. 2014a. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. *Language Documentation and Conservation* 8, 135–157
- Mosel, Ulrike (ed.). 2014b. *Kehaa*. Shellfish. Kiel: ISFAS Allgemeine Sprachwissenschaft, Universität Kiel.
- Mosel, Ulrike. (ed.). 2014c. *Amaa moon bara otei vaa Teapu*. The life and work of Teop women and men. Kiel: ISFAS Allgemeine Sprachwissenschaft, Universität Kiel.

- Mosel, Ulrike, Mahaka, Mark, Enoch Horai Magum, Joyce Maion, Naphtali Maion, Ruth Siimaa Rigamu, Ruth Saovana Spriggs, and Jeremiah Vaabero, Marcia Schwartz and Yvonne Thiesen. 2010. *Ainu. The Teop-English dictionary of house building*. Kiel: Seminar für Allgemeine und Vergleichende Sprachwissenschaft, Universität Kiel.
- Rhodes, Richard A., Leanore C. Grenoble & Anna Berge. 2006. *Adequacy of documentation: A preliminary report to the CELP*. ms.
- Schiborr, Nils N., Stefan Schnell & Hanna Thiele. 2018. *RefIND – Referent Indexing in Natural-language Discourse: Annotation guidelines (v1.1)*. Bamberg / Melbourne: University of Bamberg / University of Melbourne. (<https://www.uni-bamberg.de/fileadmin/aspra/misc/RefIND-guidelines-v1.1.pdf>).
- Schnell, Stefan. Under revision. Revisiting information management in intransitive subjects. ms.
- Schnell, Stefan & Danielle Barth. 2018. Discourse motivations for pronominal and zero objects in Vera'a. *Language Variation & Change* 30(1). 51–81. doi: 10.1017/S0954394518000054.
- Schnell, Stefan, Nils N. Schiborr, and Geoffrey Haig. 2018. Is intransitive subject the preferred role for introducing new referents? Evidence from corpus-based typology. Paper presented at the workshop Comparative corpus linguistics: New perspectives and applications, convened by Natalia Levshina & Steve Moran, SLE31, Tallinn, September 1, 2018.
- Seddhouse, Paul. 2013. Conversation analysis. In Robert Bayley, Richard Cameron & Ceil Lucas (eds.). *The Oxford handbook of sociolinguistics*, 91–110. Oxford: Oxford University Press.
- Seifart, Frank, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nivja H. de Jong & Balthasar Bickel. 2018. Nouns slow down speech: evidence from structurally and culturally diverse languages. *PNAS* 115(22). 5720–5725. <https://doi.org/10.1073/pnas.1800708115>
- Slobin, Dan. 2004. The many ways to search for a frog: Linguistic typology and the expression of motion events. In Sven Strömquist & Ludo Verhoeven (eds.), *Relating events in narrative: Typological and contextual perspectives*, vol. 2, 219–257. Mahwah, NJ: Lawrence Erlbaum.
- Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association (ELRA), 3940–3947. <http://www.lrec-conf.org/proceedings/lrec2014/summaries/1176.html>

Stefan Schnell
 stefan.schnell@uni-bamberg.de
 orcid.org/0000-0003-2036-2263

Reflections on documenting the lexicon

Keren Rice
University of Toronto

The lexicon presents unique challenges in language documentation. This reflection reviews some of those challenges, focusing on two major areas, what I have learned over time about what is important to document and the creation of dictionaries. Throughout I stress the value of considering the lexicon broadly, and, in the situation that linguists are involved, of working closely with speakers and community members in all stages of decision making, from what to document to how to spell, to how to represent meanings. N. Scott Momaday writes of words as medicine, and this is important to keep in mind in lexical documentation—one is engaging with worldview. The responsibility then of documenting the lexicon is large, and the stakes are high, given how words give deep insight into ways of being.

1. Introduction¹ The lexicon presents a challenge in language documentation, as Haviland (2006: 129) writes in his seminal paper on documenting lexical knowledge: “In the Boasian trilogy for language description of grammar, wordlist, and text, it is surely the dictionary whose compilation is most daunting. The process begins with a learner’s first encounters with a language, and it ends, seemingly, never. Worse, it is an endeavor fraught with doubt, centrally about when enough is enough both for the whole - ...- but also for any single putative dictionary entry, given the apparent endless variety of nuance and scope for words and forms, not to mention the idiosyncrasies of compounds or derived expressions...” Frawley, Hill, and Munro (2002: 1) write of how a dictionary project goes on and on, “expanding from a modest list of words and glosses to something like a cultural encyclopedia.” In this reflection, I identify some of the challenges in documenting the lexicon. My choices are idiosyncratic, reflecting my interests, but are, I suspect, of broader interest as well. In what follows I first reflect on documenting the lexicon and then on the generally expected output of documentation of the lexicon, dictionaries.

¹Many thanks to two anonymous reviewers for their comments. Thank you too to the editors for asking me to write on a topic that took me by surprise, and to Nikolaus Himmelmann for stimulating us to think carefully about what our goals are.

2. Documenting the lexicon A useful starting point in addressing lexical documentation is the description of communicative events laid out in Himmelmann (1998). He identifies a cline of event types based on planning, ranging from unplanned to planned: exclamative (ouch!, fire!), directive ('scalpel', greetings, small talk), conversational (chat, discussion, interview), monological (narrative, description, speech, formal address), and ritual (litany). While Himmelmann has little to say directly about the lexicon, the very division into different event types suggests the challenges of documenting the lexicon—there is certainly vocabulary that cross-cuts event types, but there are lexical items that are likely to occur in one type and not in others. Thus, in order to obtain a lexicon that is both broad and deep, a considerable corpus must be developed.

In an ideal world, documentation would cover all the communicative event types, from exclamatives to ritual speech, and other uses of speech that might be found. Lexical documentation would include vocabulary gleaned from different communicative event types with different participants, and from a range of semantic domains, approaching, as Frawley, Hill, and Munro (2002:1) write, a cultural encyclopedia. All material would be audio recorded, with much video recorded, and then material would be drawn from those recordings, supplemented with additional material gathered through other methods, as needed.

This is, of course, the work of lifetimes. What should be prioritized? What can wait? There are no simple, straightforward answers to these questions but I will reflect on some of what I have learned from involvement with such work over some time period.

2.1 Documenting the 'unusual' I begin at a level of speech that I did not consider when I began to do fieldwork, what I will call interjections. I have a grandchild who is being raised bilingually, in English and French, with more exposure to French than to English. What does he take to in English? He picks up on words like 'wow', 'oopsy-daisy', 'oops', 'uh-oh', 'ouch', and so on very quickly, and he likes to use them, and to reflect on the difference between words like 'oops' and 'uh-oh.' I have not heard him using this type of French word in English where an English one is appropriate, although he frequently uses French nouns and verbs when speaking with monolingual English speakers. There is something special about this part of the lexicon, and, while I had not thought to document it, although I heard it, watching my grandson acquiring two languages makes me realize how important such vocabulary is—it gives some kind 'feel' to the language. Similar is baby talk, talking to pets and animals, sounds that animals make, colloquial expressions, directives, and other communicative types where words or set phrases are likely an appropriate unit. Meanings might be difficult to express in another language, but this type of vocabulary is special, and knowledge of such words can be taken as a sign of cultural knowledge.

2.2 Etymologies My next point is something that I have come to understand from working with Elders from several Canadian communities over the years, namely trying to understand what some think of as the 'true meaning' of a word. Some question the value of studying etymology. Mosel (2011:349), for instance, in writing about lexicography, suggests that although many people are interested in the history of languages, documentation of etymology should be postponed, with documentation of the living language taking priority. Whaley (2011:343) asks if thinking about etymology might be a case of "loving the language more than loving its speakers." Kroskrity (2015:151) notes a conflict in that seeking to understand the etymology of words can be

interpreted as privileging the past rather than the present. Nevertheless, in my experience an understanding of what words mean, where they come from, and their internal parts is valued by many involved in understanding what the values of their society are, and seeking to rebuild strengths that have been challenged through colonization. This is not to privilege the past, but to dream for the future. Perhaps the people who are keen about this are the philosophers and historians of their societies, but their desire to understand those values is important to them. For instance, Goulet and Goulet (2014:60), in their book on Nehinuw concepts and Indigenous pedagogies, discuss life force, writing that Nehinuw illustrates the interactive, dynamic process of causal forces that is deeply embedded in Nehinuw traditional culture. As one example, they give the root/stem waso ‘she, he, it shines,’ noting that it occurs in words like ‘sun’ and ‘stars,’ and also in the word awasis ‘child,’ or ‘the little being that shines,’ with children epitomizing the “light, sparkle, and vibrancy of life” in traditional Nehinuw culture. Drabek (2018) writes of how understanding the meanings of words of her ancestral language Kodiak Alutiiq help her to see the world in a different way. She considers the word, -imaq ‘sea, ocean,’ also used for ‘a liquid contained inside’ and contents. She writes how this word inspires her—*imartuq* ‘it is full,’ *imaituq* ‘it is empty,’ *imasuugtua* ‘I feel depressed, or sad, I am downhearted, I have a sinking feeling of foreboding,’ or, more literally, ‘I am searching for my contents.’ Kroskrity (2015:151), focusing on designing a dictionary, discusses a Tewa verb that gives a sense of scarcity, at least from an historical perspective. He notes that the verb is restricted in the objects it can refer to, with those objects being things that were precious or vital to the well-being of the community. While much has changed, understanding the lexical semantics of this verb gives clues into values and worldview. Thus, understanding etymologies and lexical semantics can lead to insights that take one far beyond the language, to understanding values. This is, for some, not something to be disregarded as the etymology and meaning provide links between past, present, and, hopefully, future.

2.3 New vocabulary Another controversial area is new vocabulary, both loanwords and newly created vocabulary. Loanwords may or may not be recognized as such by speakers, depending on how much they are integrated into the language, broader knowledge of the speaker, and so on. New words are created to represent new things. Are such words included in documentation? The answer could be yes or no—although the ‘ideal’ documentation project might include new words as well as information about the process by which the language captures ‘new’ ideas, whether those new words are included in a lexicon depends on what is privileged in the particular context. It would be difficult to document without encountering loanwords and new words if speakers talk about a wide variety of topics, different communicative events with different participants are recorded, and so on. But if the focus is on tradition, such words might not occur.

The Alberta Elders’ Cree Dictionary, compiled by Cree elders (LeClaire and Cardinal 1998), comments on new words: “A dictionary of this sort is not just a collection of words and their meanings, but represents something of what the community it serves requires. Hence, we have incorporated suggestions from a wide variety of Alberta bands for making this dictionary more usable for their members. [W]hat words we thought helpful, though not yet accepted widely by Cree speakers, or words that reflect recent English influence, or idiomatic Cree that did not appear connected directly to traditional Cree usages are ... found in the supplemental” The Elders who compiled the dictionary

believed that an indication of vitality of a language was the ability of speakers to talk about what was around them.

2.4 Documentary methods It is worthwhile to very briefly consider methods of documenting the lexicon. Himmelmann's focus on communicative events can present challenges in documenting the lexicon, as much might not emerge in even a reasonably large corpus. In documenting semantic domains, there has thus been continuing emphasis on ways of documenting the lexicon that value teamwork with local and academic experts; see Evans (2012) for an anecdote on the importance of this. Many people I have worked with are keen to work on vocabulary in particular semantic domains, not just those that might be considered of relevance from a cultural perspective, but also things like types of footwear (including, in addition to words for moccasins, mukluks, and the like that are considered culturally significant, also words for running shoes, high heels, and so on). As one seeks to document the lexicon as fully as possible, awareness not just of the past, but also of the present, matters.

3. Thinking about dictionaries Ogilvie (2011:402), examining the effects of language documentation on lexicography, concludes that "the lexicographer cannot ignore the new focus on primary data; the new recognition of the importance of collaboration and involvement of the speech community in the dictionary-making process; the new concerns for accountability and ethics; the new concern for storage and accessibility of archived dictionary materials; and the new possibilities that technology brings to both the content of dictionaries and their compilation." Ogilvie writes (2011:393) that dictionaries "have begun to blur the boundaries between documentation and description," commenting on how dictionaries are repositories for primary data, including images, sound, and video. While this blurring exists, I reflect on documentation—collection of primary materials—and description in the form of dictionaries separately, now addressing dictionaries.

Perhaps the major change since Himmelmann (1998) involves developments in technology. These allow for more kinds of dictionaries: talking dictionaries, dictionaries with videos, user-driven dictionaries. I do not pursue technology but would be remiss in not noting its tremendous role in advancing lexicography. Nor do I address the many other issues that arise in the discussion of dictionaries, including the nature of the word, the challenges of lexical analysis, and the content and organization of dictionaries.

In documentation, dictionaries are often designed with language sustainability or reclamation in mind (this is not to devalue fuller dictionaries, but such dictionaries take years and having products along the way is of value). While defining goals is worthwhile in determining structure and content, goals alone do not provide easy answers for many questions. I address three, standardization, meanings, and, in a different vein, issues of control.

3.1 Standardization When I first began working with people on a dictionary, spelling seemed like a relatively unimportant issue. At that time, several years had been spent by committees working on standardization of symbols and spellings, and decisions had been reached on symbols to use, on how to spell, and on principles to follow in light of variation by and between speakers.

Starting with symbols, I became unsure why IPA vowel values have been so widely adopted in situations where speakers are familiar with values associated with English

vowels. I have been asked numerous times to put something in a spelling that English speakers can relate to more easily (e.g., instead of *tu*, write *too*, with the vowel like that in ‘*too*’). It would have been difficult for the standardization committees to change symbols, given the material that existed with spelling based on IPA vowels, but the initial decision to use IPA was perhaps ill-conceived, not taking into account the likely users of the dictionaries and the background that they had.

Work on a dictionary gave rise to other surprises, as some of the principles of standardization that had been agreed on by standardization committees became sources of concern. Rice and Saxon (2002) discuss this, as do Hinton and Weigel (2002), Mosel (2011:341), and others. Mosel suggests that while standardization is often a political matter that can be difficult to resolve, having a standardized spelling is nevertheless important in a dictionary. The standardization committees recommended a single spelling reflecting conservative usage. But I learned that community developers of a dictionary do not necessarily agree. As some I’ve worked with have said, spelling standardization might be conceived of as a western ideal that is not held in all societies. Standardized spelling might emerge over time, but need not be the starting point, as it excludes rather than includes people, privileges some over others, and makes the relationship between the oral and the written more obscure, something that was not valued—people wanted to hear the voice of the person as they read their words.

Standardized spelling has proven to be untenable in current dictionaries that grow out of community work. In a Dëne Sə́łíné Yatié dictionary (Kaulback, Catholique, Drygeese 2014:11), for instance, the editors write of varieties and choices about spelling: “The changes and this variability are problematic when one begins the process of recording these words and preserving them in a written form.... We are cognizant of the fact that there are speakers that use the K-dialect in the community and others who don’t. We are also aware that some words have been shortened but there remains a longer—some would say purer—form of the same word still in use among some speakers. In an effort to create a resource that best represents the language of the community and accounts for this variability, we have included alternate spellings and pronunciations for many words We encourage the reader to find their language within these words recognizing, of course, that not all forms of a word may be accounted for. ... The elders recognized the value of this dictionary and the integrity of the process, and contributed to it with all their hearts. ...”

Variation, I have learned, is well accepted in many places, and singling out one variety as ‘better’ may well be culturally inappropriate, at least at early stages.

3.2 Meanings and cultural concepts Much has been written on semantic fields such as traditional tools, kinship, toponyms, astronomy, cooking, ethnobotany, and government. These are important topics, and worthy of inclusion in a dictionary, and are often located in thematic dictionaries or thematic parts of a comprehensive dictionary. Semantic fields can bring surprises. Once I was working with someone on classifying words into semantic fields, and we were discussing what field the words ‘*bow*’ and ‘*arrow*’ belonged in. I assumed hunting, and was taken by surprise when the response was that they were toys. For some they would have been considered hunting, but they are no longer used for that today, and this was what mattered to the person who I was working with.

In addition to semantic fields, aspects of meaning exist that are difficult to capture in translation, but give clues into values and worldview. I illustrate with an anecdote. I did a lot of knitting when I was doing fieldwork, and people often commented to me

that I was wasting my yarn. I found this an odd, even offensive, remark—I didn't think that I was wasting it, but rather that I was using or transforming it. It took me time to understand what was going on, and it is easiest to talk about this by giving some cultural background. Rushforth and Chisholm (1991) discuss cultural persistence in the Sahtú people of Canada's Northwest Territories, where I was doing fieldwork, introducing a Sahtú concept of *séodjit'é*, or what a culturally ideal person is like. Such a person is described as shy, humble, non-imposing, careful, caring, reasonable, reserved, controlled, polite, industrious, generous, restrained. This concept is important for understanding the nature of the lexicon. There are numerous pairs of verb stems, both of which translate roughly the same. For instance, the stems *-ta* and *-?e* both mean do something with the foot. There are, however, subtle differences between these that are hard to translate into English, and these differences relate to *séodjit'é*. Returning to the anecdote, I learned that what was translated into English as 'waste' is not negatively valued in the language, while it is in English—to put this another way, *séodjit'é* is positive but the reverse is not negative. How to address this in a dictionary—I still don't know the answer—words like 'gentle' vs. 'rough', 'slowly' vs. 'quickly' are used but they do not capture the essence of the difference. I do know that understanding the difference between these words provides a clue to understanding values of deep cultural importance, something that is expressed many aspects of the language as well as in ways of living.

3.3 Control I end with a discussion of control. While this does not have to do with the lexicon per se, it is an important topic as societies grapple with issues of privilege and power. The lexicon is a topic that, in a project involving an outside linguist, requires close collaboration with speakers. In working on dictionaries in documentation, work that blurs the lines between academic and community is increasingly understood to be important in many settings. A dictionary is, ultimately, a product designed for use by a community. There probably is no one right starting point for a dictionary beyond what people are interested in. It might be that the interest is in cultural traditions. But it might be that it is in naming new things. A dictionary of the type people are familiar with for languages like English, French, and Spanish may be a vision, but may be a barrier to dictionary creation. In the end, dictionaries are lists of words and phrases organized in some way so that words can be found, with, minimally, information about pronunciation, meaning, and use. Starting with expectations about what a dictionary should be can produce dictionaries that are rejected by a community; working collaboratively, trying to understand the vision for a dictionary at a particular time, trying to let go of pre-conceived notions of what a dictionary must be can lead to a very different type of dictionary than one could have imagined at the start.

Words matter. The Kiowa novelist, poet, and essayist N. Scott Momaday (1968:89) makes this clear in the following quote: "Words were medicine; they were magic and invisible. They came from nothing into sound and meaning. They were beyond price; they could neither be bought or sold ..." As a reviewer remarks, in engaging with the lexicon, one is engaging with worldview and ways of being. Thus, the stakes of lexical documentation and the resulting products can be very high, and all engaged must consider seriously the import of what they do, while at the same time delighting in how much they learn, both expected and unexpected.


References, plus some works that provided inspiration

- Drabek, Alisha. 2018. Echoes from my Kodiak Alutiiq ancestors. (<https://www.humansandnature.org/echoes-from-my-kodiak-alutiiq-ancestors>) (Accessed 25 April 2018)
- Evans, Nicholas. 2012. Anything can happen: The verb lexicon and interdisciplinary fieldwork. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 183–208. Oxford: Oxford University Press.
- Frawley, William, Kenneth C. Hill & Pamela Munro. 2002. Making a dictionary: Ten issues. In William Frawley, Kenneth C. Hill & Pamela Munro (eds), *Making dictionaries: Preserving Indigenous languages of the Americas*, 1–22. Berkeley: University of California Press.
- Genee, Inge & Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation & Conservation* 12. 274–314.
- Goulet, Linda M. & Keith Goulet. 2014. *Teaching each other: Nehinuw concepts and Indigenous pedagogies*. Vancouver: UBC Press.
- Grenoble, Lenore & Simone S. Whitecloud. 2014. Conflicting goals, ideologies and beliefs in the field. In Peter K. Austin & Julia Sallabank (eds), *Beliefs and ideologies in language endangerment, documentation and revitalisation. Proceedings of the British Academy* 199, 339–356. Oxford: Oxford University Press.
- Haviland, John. 2006. Documenting lexical knowledge. In Gippert, Jost, Nikolaus Himmelmann & Ulrike Mosel (eds), *Essentials of language documentation*, 129–161. Berlin: Mouton de Gruyter.
- Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Hinton, Leanne & William Weigel. 2002. A dictionary for whom? Tensions between academic and nonacademic functions of bilingual dictionaries. In William Frawley, Kenneth C. Hill & Pamela Munro (eds.) *Making dictionaries: Preserving Indigenous languages of the Americas*, 155–170. Berkeley: University of California Press.
- Kaulback, Brent, Bertha Catholique & Dennis Drygeese (eds). 2014. *Dënë Sṓliné Yatié. ?ereht'ís. ʔuskēlk'e t'iné Yatié. Chipewyan Dictionary*. South Slave Divisional Education Council.
- Kroskrity, Paul. 2015. Designing a dictionary for an endangered language community: Lexicographical deliberations, language ideological clarifications. *Language Documentation & Conservation* 9. 140–157.
- LeClair, Nancy & George Cardinal, edited by Earl Waugh. 1998. *Alberta Elders' Cree dictionary / Alperta ohci kehtehayak nehiyaw otwestamâkewasinahikan*. Edmonton: University of Alberta Press.
- Momaday, N. Scott. 1968. *House made of dawn*. New York: Harper & Row.
- Mosel, Ulrike. 2011. Lexicography in endangered language communities. In Peter K. Austin & Julia Sallabank (eds.), *Cambridge handbook of endangered languages*, 337–353. Cambridge: Cambridge University Press.
- Ogilvie, Sarah. 2011. Linguistics, lexicography, and the revitalization of endangered languages. *International Journal of Lexicography* 24(4). 389–404.

- Rice, Keren & Leslie Saxon. 2002. Issues of standardization and community in Aboriginal language lexicography. In William Frawley, Kenneth C. Hill & Pamela Munro (eds.), *Making dictionaries: Preserving Indigenous languages of the Americas*, 125–154. Berkeley: University of California Press.
- Rushforth, Scott & James Chisholm. 1991. *Cultural persistence: continuity in meaning and moral responsibility among the Bearlake Athapaskans*. Tucson: University of Arizona Press.
- Whaley, Lindsay. 2011. Some ways to endanger an endangered language project. *Language and Education* 25(4). 339–348.

Keren Rice

rice@chass.utoronto.ca

 orcid.org/0000-0002-8112-8908

Reflections on linguistic analysis in documentary linguistics

Bradley McDonnell
University of Hawai'i at Mānoa

This article reflects on the role of analysis in language documentation since Himmelmann (1998). It presents some of the criticism that Himmelmann's notion of analysis faced and how he responded (Himmelmann 2012). However, analysis in this context rarely refers to analysis alone, but the term includes the larger research goals and research questions. This study, then, situates the research goals, research questions and analyses that I have employed in my research on Besemah on a cline from *facilitative* to *restrictive* in terms of the diversity and spontaneity of the (archival) record that is produced, building upon Himmelmann's (2012) conceptual basis for distinguishing documentation and description. It does so through two case studies in Besemah, one with a highly facilitative research goal, question, and analysis and another with a highly restrictive research goal, question, and analysis.

1. Introduction¹ The role of analysis in language documentation—or the perceived lack thereof—has been one of the most contentious issues in language documentation since Himmelmann (1998). This is largely a reaction to Himmelmann's sharp distinction between language documentation and language description as two separate fields of inquiry. In my reflection, I very briefly review some of the criticism that Himmelmann's original proposal faced (e.g., Rhodes et al. 2006, Evans 2008) and how he has responded

¹I would like to thank the participants from the village of Karang Tanding who took part in the Besemah conversations and in the word-stress experiment. I would especially like to thank Sarkani, Hendi, and Sutarso who helped transcribe and annotate the conversations. I am grateful to Asfan Fikri Sanaf and Kencana Dewi for kindly allowing me stay at their home in Karang Tanding, and the Language Institute at Sriwiaya University and the Center for Culture and Language Studies at Atma Jaya Catholic University of Indonesia for sponsoring this research. It was funded by a Fulbright scholarship, Blakemore Freeman Fellowship Language Grant, and Fulbright-Hays Doctoral Dissertation Research Abroad scholarship. I am grateful to the Ministry of Research and Technology in Indonesia for allowing me to conduct this research on Besemah. Finally, I would like to thank Gary Holton, Andrea Berez-Kroeker, and two anonymous reviewers for their comments on an earlier draft of this paper.

(Himmelmann 2012). Through discussion of linguistic analysis in these papers, it has become clear that this term represents more than analysis itself but is almost always tied to larger research goals and research questions as well as issues of data collection (i.e., how documentary linguists decide on the data to be gathered and go about gathering them and—often based upon the analysis—go about annotating them). Based on these discussions, I show from my research on Besemah (iso 693-3: pse), a Malayic language of southwest Sumatra, that the types of analyses that I chose to utilize as well as their associated research goals and questions had a significant effect on the documentation of Besemah. More specifically, I illustrate how the analyses that I chose to employ, based on different research goals and questions, had consequences for (i) the types of (raw) data that were archived, and (ii) the extent to which these data were annotated in the process of analyzing them.

Building upon Himmelmann's (2012) conceptual basis for distinguishing documentation and description, I situate the research goals, research questions and analyses that I have employed in my research on Besemah on a cline from *facilitative* to *restrictive* in terms of the diversity and spontaneity of the (archival) record that is produced.² Maximally facilitative research goals, research questions and analyses allow for the types of data collection and annotation that result in a documentation that is spontaneous (i.e., not constrained by a researcher's task), diverse (i.e., encompassing many different types of speech events), and richly annotated (i.e., with information on all types of linguistic and non-linguistic factors). The documentation is enriched on various levels based primarily upon the analytical path one chooses to follow, but it does not put many constraints on the data that are collected. Maximally restrictive research goals, research questions and analyses allow for the creation of a dataset that is controlled in such a way as to avoid confounding variables and allow for a more straightforward analysis that better answers specific research questions that need such control. Facilitative research goals, research questions and analyses arguably have less researcher bias in data collection, and the resulting documentation may prove to be more useful to a wider range of audiences in the long term. It allows for a documentation that has the best chance to answer questions that we have not yet thought to pose. Restrictive research goals, research questions and analyses, on the other hand, purposely bias the data collection, and the resulting archival record are less likely to create a dataset that will be widely utilized in the long-term or used to answer questions that have not yet been asked.

From this perspective, it seems to me that descriptive linguistics tends to set research goals, pose research questions and draw on analyses that would be located somewhere in the middle of this cline, and researchers now and in the future may benefit from incorporating different types of research goals, research questions and analyses that fall on the extreme ends of this cline (i.e., those that are more facilitative and to a lesser extent those that are more restrictive). That is, it is my impression that descriptive linguistics ask research questions and employ analyses that draw heavily on semi-spontaneous collections of staged narratives (i.e., narratives told for the purpose of documentation) and artificial tasks (e.g., Pear Story (Chafe 1987), Frog Story (Berman 1994) or SCOPIC (Barth & Evans 2017)) on the one hand, and targeted elicitations based upon manipulated or invented examples for the purpose of filling in a paradigm or obtaining grammaticality judgments on the other hand. Often facilitative research questions and analyses that

²Spontaneity here is similar to what Himmelmann (2012) refers to as "direct input from native speaker", which is divided into two groups: "data based on observable linguistic behavior" and "data based on metalinguistic skills" (199).

result in the collection and annotation of everyday conversations are often marginal if represented at all in a documentation. Presumably, this is because interactional data is much more difficult to process and analyze. Likewise, carefully controlled experiments that test well-defined hypotheses, which fall on the restrictive end of the cline, are also quite uncommon due to a number of factors related to training and the practicalities associated with running such experiments.

The next section outlines controversies surrounding analysis since Himmelmann (1998), and then section 3 presents two examples from my research on Besemah, one facilitative, which involves the annotation of conversational data, and one restrictive, which involves an experiment on word stress that collected a controlled dataset.

2. Controversies over analysis Himmelmann's (1998) proposal to create a sharp distinction between documentation and description—while it was met with much enthusiasm (see Austin 2016)—faced both skepticism and criticism (Evans 2008, Rhodes et al. 2006, Chelliah & de Reuse 2011, Woodbury 2011), and much of this criticism addressed issues surrounding the role of analysis in language documentation. For example, Rhodes et al. (2006), in an unpublished report to the Linguistics Society of America's Committee on Endangered Languages and their Preservation (CELP), responded to Himmelmann's distinction between language documentation and language description by emphasizing the importance of a systematic analysis of a language. They contend that there is an important *accounting function* of analysis, which holds the view that the systematic analysis that one does during the production of descriptive materials (e.g., a reference grammar) is essential for the documentary linguist to know what has been documented and what still needs to be done. Their view is summarized by the following quotation:

Himmelmann (1998) has argued persuasively that documentation is distinct from what he calls description, i.e., linguistic analysis. We think this is seriously mistaken. In order to know how far along one has come in documenting a language one must be able to measure how far there is to go. A crucial part of that measurement is found in the accounting function of analysis. How do we know when we've gotten all the phonology? When we've done the phonological analysis, and our non-directed elicitation isn't producing any new phonology. How do we know when we've gotten all the morphology? When we've done the morphological analysis, and our non-directed elicitation isn't producing anything [sic] new forms, and — crucially in inflected languages — when we elicited all the implicit inflected forms that haven't happened to come up in non-directly [sic] elicitation. (3)

Evan's (2008) review of Gippert et al. (2006)—but in reference to Himmelmann (2006) more specifically—follows up on the point made by Rhodes et al., where he notes various examples for which carefully controlled elicitation was needed to understand key concepts about the phonology and the grammar of various languages. He also points out the importance of the thorough analyses that result from a reference grammar where he criticizes Himmelmann's approach to language description: "To see this as mere formulation and organization is to grossly underestimate the nature of the analytic challenge" (348). Both Evans (2008) and Rhodes et al. (2006) maintain the position that the systematic analysis that is found in language description is not ancillary to language documentation, but is a crucial element of it.

In addition to these explicit criticisms of Himmelmann's conception of language documentation, there is a general misunderstanding that language documentation concerns the amassing of data without any analysis or that language documentation even opposes analysis in some way (Himmelmann 2012: 1). However, analysis has always been a part of Himmelmann's conception of language documentation. For example, Himmelmann (1998) proposes that there be a mutual dependency between analytical frameworks (e.g., sociolinguistic and anthropological approaches to language, phonetics, corpus linguistics, etc.) and the documentation, wherein the process of collection and presentation of a language documentation is significantly influenced by the analytical framework. On a more practical level, Himmelmann (1998) addresses the need for analysis in the transcription, translation, and commentary in the documentation of speech events, which allows the documentation to be accessible to a wide range of audiences. Therefore, the crux of the debate over analysis is not whether documentary linguists should incorporate analysis in their documentations but to what extent can documentary linguistics separate out the activities associated with language documentation from the systematic analysis of particular phenomena found in descriptive linguistics (see also Chelliah & de Reuse 2011).

In a follow up to Himmelmann (1998, 2006), Himmelmann (2012) addresses these issues more explicitly and from both theoretical and practical perspectives. In theory, he proposes a model that distinguishes data types based on input from native speakers: data based on observable linguistic behavior (e.g., recording of a conversation) and data based on metalinguistic skills (e.g., elicitation). These data types intersect with three stages of data processing: (i) processing raw data (e.g., recording audio/video recording), (ii) processing primary data (e.g., transcription and translation of recording), and (iii) developing structural data (e.g., descriptive generalizations, interlinear glosses). For Himmelmann this model helps in the delineation of activities concerned with documentation and those concerned with description, which is summarized in the following quotation:

Documentary linguistics ... is primarily concerned with raw and primary data and their interrelationships, including issues such as the best ways for capturing and archiving raw data, transcription, native speaker translation, etc. Descriptive linguistics ... deals with primary and structural data and their interrelationships ... Primary data ... thus have a dual role, functioning as a kind of hinge between raw and structural data. They are the result of preparing raw data for further analysis (documentation), and they serve as input for analytical generalizations (description) (2012: 199).

In practice, however, Himmelmann (2012) recognizes that documentary linguists are not necessarily going to create neat distinctions between these types of activities. Thus, he provides a pragmatic resolution: "Do what is pragmatically feasible in terms of the wishes and needs of the speech community and in terms of your own specific skills, needs, and interests" (201). This resolution provides a lot of freedom for the linguist to document what they are best trained to do and satisfies the needs and desires of the community. It also fits into a larger shift in language documentation to individualized approaches that are tailored to the social, cultural, and political contexts in which they occur (Austin 2016). This resolution, while providing the documentary linguist more freedom in terms of their choice of research goals, research questions and analyses to be employed, raises important issues regarding the resulting documentation and its usefulness, especially in

its ability to be useful to different users in the long-term and answer questions that no one has thought to pose. *How useful is the documentation for a broad audience? How useful will it be in the long-term? How likely can it be used to answer other yet to be posed questions or be used for other purposes?*

These questions harken back to the original reason that Himmelmann (1998) proposed to separate documentation from description: data collections and their annotations tended to be limited to serve descriptive goals and lacked long-term usefulness to a broad audience. In my own research, I have found that beyond Himmelmann's pragmatic resolution, it is also important to reflect on how given research goals, research questions and/or analyses *facilitate* a documentation that is long-lasting and potentially useful to a wide range of audiences or *restricts* the dataset that is intended to answer only current research questions.

3. Different approaches This section briefly reflects on the effects of adopting research goals, research questions, and analyses that fall on either end of the facilitative-restrictive cline. I demonstrate this with two very different studies that I employed in my research on Besemah. The first study is highly facilitative and concerns voice selection in everyday conversation in Besemah. This study shows that while the research question is focused, (i) my larger research goal to understand structures that arise in the course of everyday conversations allowed me to collect data that is broadly useful for various purposes, and (ii) the research question required intimate knowledge of social, cultural, and interactional contexts of the everyday conversations in the corpus, which ultimately resulted in rich annotation of these speech events.

The second study is highly restrictive and concerns the status of word-level stress in Besemah. It shows how the data collection is restricted because it is directly tied to both the research question and subsequent analysis that crucially requires control of confounding variables. It is important to note that while I think both highly facilitative and highly restrictive research goals and analyses are important to answer different research questions, they are not equally important for language documentations. Highly facilitative research goals, research questions, and analyses are much more important for a language documentation, and highly restrictive research goals, research questions, and analyses may serve an important supportive role for a language documentation. The importance of this supportive role is difficult to predict in the long term as we cannot know what questions will be important for future users of the documentation, but as we will see they are clearly important to answer current questions.

3.1 Highly facilitative analysis While the inclusion of conversation in language documentations have been generally advocated for (Himmelmann 2004, Sugita 2011, Childs et al. 2014, Austin 2016), there has been little emphasis on how best to collect or analyze conversation in a language documentation context (McDonnell forthcoming). Field linguists have long recognized the difficulty of working with conversational data as narrative data is much easier to collect, analyze, and exemplify in writing.

However, everyday conversation is ubiquitous, and its documentation is vital (Levinson 2006, Childs et al. 2014). It is both useful for broad audiences and in the long-term has potential to answer questions that we have not yet thought to pose. Thus, research goals that seek to answer particular research questions about everyday conversation are highly facilitative because data collection is typically not tied to any particular research question and the documentation is not constrained by it.

The study presented in this section is a case in point. The conversations were collected before I posed any particular research question. Equally important is the fact that the analysis of everyday conversation, which this study required, allowed for the creation of rich annotations at many different levels, including (i) extensive glossing and additional semantic and morphosyntactic annotations and (ii) annotations of sociocultural knowledge and interactional practices. See McDonnell (forthcoming) for further description of these types of annotation, especially those in (ii) above.

Voice selection study My study of the *symmetrical voice* system (i.e., a voice system with two or more transitive voices, neither of which is derived from the other) in Besemah (McDonnell 2016) exemplifies these points concerning facilitative research goals, research questions, and analysis well. In this study, I was interested in answering a straightforward research question: *At any given point in a conversation, what led to the use of one voice over the other?* In Besemah, this was particularly interesting because each of the two voices (i.e., the agentive voice and the patientive voice) were quite common in conversation; agentive voice occurred approximately 60% and patientive voice approximately 40% of the time. I chose to answer this question using methodologies from Usage-based linguistics (Bybee & Beckner 2009), Interactional Linguistics (Selting & Couper-Kuhlen 2001) and quantitative corpus linguistics (Gries 2017).

In order to answer this research question, I drew on a handful of recordings of everyday conversations that I had collected earlier, based upon my larger research goals to understand structures that arise in the course of everyday conversations. These recordings had been transcribed, translated into Indonesian and English and received some glossing. Once I had a clear research question in mind, I cleaned up the transcription, translations, and glosses and finished glossing the remainder of the recordings. Most importantly, with the help of several Besemah language consultants, I provided commentary about the larger context in which the symmetrical voice alternations occurred as well as the semantic and syntactic information about predicates and their arguments. These annotations were included in a notes tier in ELAN. This documentation was, in turn, further coded for quantitative analysis, including several morphosyntactic properties (e.g., transitivity, presence of causative/applicative suffix, person-number of arguments) and discourse properties (e.g., information status, specificity of arguments, topic continuity) based upon the detailed documentation (i.e., transcriptions, translations, glosses, and commentaries). For a detailed discussion of this process see McDonnell (2016: 201-226).

While it is unclear how useful the detailed coding would be to a broader audience, the annotation created during fieldwork is widely accessible and broadly useful. Besides transcriptions, translations, and glosses, it provides commentary on the grammar of the language, but more importantly it provides commentary that contains important cultural information, broader social context, and information about speakers and referents. For example, understanding the information status of referents, whether or not they occurred in a symmetrical voice construction, was critical to my study. However, Besemah speakers rarely refer to someone by name once that person has children, instead they commonly refer to them using *bapang* ‘father’ or *endung* ‘mother’ and the name of their child (e.g., *endung Refki* ‘Refki’s mother’) or using a kinship term *mamang* ‘uncle’ or *bibik* ‘auntie’ and their eldest child’s name (e.g., *bibik Refki* literally means ‘Refki’s auntie’ or ‘auntie Refki’ but is commonly used to refer to ‘Refki’s mother’).³ This use

³This phenomenon is known as *teknonomy* in Anthropology.

of the kinship term and eldest child's name created much ambiguity for someone who does not have intimate knowledge of the people discussed in the recordings and the sociocultural contexts in which these reference terms are used. In many cases, the same individual was referenced in many different ways, using different kinship terms. Thus, in annotating these conversations, I consistently asked Besemah language consultants about such referents, and included information on who the speaker was making reference to and why the speaker used the particular kinship term they did. This information, which is part of a larger category of annotation that Schultze-Berndt (2006) refers to as contextual commentary, helps future users of this documentation interpret the conversation.

The analysis of symmetrical voice then facilitated a rich documentation not just of symmetrical voice or even Besemah grammar but of various aspects of everyday conversations in Besemah. It led to me to ask questions about the conversations themselves that I would have likely not asked otherwise, and it forced me to understand the conversational contexts much more deeply. Most importantly, the annotation that my analysis produced has the best chance to be useful for the long-term to answer questions that have not been posed because they serve an accessibility function that helps future users of the documentation to understand it.

3.2 Highly restrictive analysis While some have advocated that elicitation take on a more experimental approach (e.g., what (Yu 2014) refers to as “an experimental state of mind”), most current elicitation practices in descriptive linguistics differ from experiments in the follow ways:

1. Stimuli and design are set for all participants in experiments but elicitation tasks are often adaptive, where linguists react to the responses of the language consultant;
2. Experiments set out to test specific hypotheses while elicitation often involves a process of honing different hypotheses until there is something testable;
3. Experiments typically involve statistical analysis and elicitation rarely if ever does.⁴

Thus, experiments represent more restrictive research goals, research questions, and analyses because the data collection is constrained by the analysis and experimental design, which requires control; data are collected with specific research questions and necessary controls in mind. The study I present in this section on Besemah word stress exemplifies these restrictive research goals, research questions, and analyses well. It demonstrates how experimentation is necessary to address questions that cannot be answered by elicitation alone but critically rely on control and a subsequent statistical analysis. It also demonstrates that the dataset that results from such a study is restricted in terms of its limited usefulness beyond the present study.

Word stress study Recent studies have shown that the analysis of word-level stress is not at all straightforward (Gordon 2014), and in the description of many languages word-level stress and phrase-level prominence appears to have been conflated. That is, language descriptions appear to commonly describe word-level stress based upon words

⁴In some cases, the distinction between elicitation and experimentation is less clear. For example, the frog story was originally collected in a controlled way and with a particular hypothesis in mind (Berman 1994). Nowadays, it is my impression that the frog story is collected as a way to elicit a story with relative ease. The point here is that there are distinctions between prototypical elicitation and experimentation.

uttered in isolation, which means that the word in and of itself is a complete utterance. Thus, word- and utterance-level prominences are conflated, which make it impossible to tell whether the prominence is attributed to the word, the utterance, or a combination of the two. Recent studies have shown this to be particularly prevalent in the languages of Indonesia (van Zanten & van Heuven 1998, 2004, van Zanten & Goedemans 2007, van Heuven et al. 2008, Maskikit-Essed & Gussenhoven 2016).

To answer this question for Besemah, I designed an experiment that carefully controlled for various word- and utterance-level factors and subsequently conducted a quantitative analysis of the acoustic properties of these words. Details of the experiment and analysis are described in McDonnell & Turnbull (2018). The basic idea is that if I control for word- and utterance-level factors, then I can attribute a result (e.g., differences in pitch, intensity, spectral balance, etc.) to either prominence in the word or the utterance. The types of control imposed on the recordings for word-level factors include (i) matched vowels within the words (e.g., /pipis/ ‘pulverize’, /tatap/ ‘touch’) to control for intrinsic acoustic properties of vowels (i.e., low vowels have intrinsically higher intensity) and (ii) balance in the weight of syllables within the word (e.g., words with two light syllables or two heavy syllables, a light followed by a heavy syllable or a heavy followed by a light syllable) to control for the possibility of stress being attracted to heavy syllables. Utterance-level controls include (i) varying the position of the target word within the utterance (e.g., the word appears in a phrase medial position or a phrase final position) in order to understand any interactions with utterance-final pitch excursions (i.e., boundary tones) and (ii) varying the information status of the target word (e.g., whether the word is ‘in-focus’ or ‘out-of-focus’).

What is important to note here is that the archived dataset that results from this highly restrictive analysis provides crucial evidence for word-level stress. Such an analysis would not be possible if it were based upon elicitation under a descriptive framework. The analysis is, however, only as good as the design in which it was collected and the dataset underlying the analysis is unlikely to be re-purposed in any significant way. They would likely only be useful for the purposes of reanalysis (e.g., based on new acoustic measurements or a new quantitative analysis).


4. Conclusion In this reflection, I have shown how the different research goals, research questions, and analyses affected the archival record of Besemah. I show that these can be highly facilitative, which allow for rich documentation both in terms of the raw data that is collected and the annotation created, or highly restrictive, which allows us to answer specific (hypotheses testing) questions that cannot be answer through elicitation. From this reflection, I hope to encourage researchers interested in language documentation to consider their research goals, research questions and analyses in terms of the resulting archival record of the language, even to the point that researchers could take a more radical approach and alter the research goals they set, the questions they pose, and the types of analyses they employ to ensure a rich archival record of the language.

References

- Austin, Peter K. 2016. Language documentation 20 years on. In Luna Filipović & Martin Pütz (eds.), *Endangered languages and languages in danger: Issues of documentation, policy, and language rights* (IMPACT: Studies in Language and Society 42), 147–170. Amsterdam: John Benjamins Publishing Company. doi:10.1075/impact.42.07aus.
- Barth, Danielle & Nicholas Evans. 2017. *Social Cognition Parallax Interview Corpus (SCOPIC)* (Language Documentation & Conservation Special Publication 12). <http://hdl.handle.net/10125/24739>.
- Berman, Ruth Aronson. 1994. *Relating events in narrative: a crosslinguistic developmental study*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Bybee, Joan L. & Clay Beckner. 2009. Usage-based theory. In Bernd Heine & Heiko Narrog (eds.), *The oxford handbook of linguistic analysis*, Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199544004.013.0032.
- Chafe, Wallace L. 1987. Cognitive constraints on information flow. In Russell S Tomlin (ed.), *Coherence and grounding in discourse* (Typological Studies in Language 11), 21–51. Amsterdam: John Benjamins.
- Chelliah, Shobhana L & Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.
- Childs, Tucker, Jeff Good & Alice Mitchell. 2014. Beyond the Ancestral Code: Towards a Model for Sociolinguistic Language Documentation. *Language Documentation & Conservation* 8. 168–191.
- Evans, Nicholas. 2008. Review of Essentials of Language Documentation. *Language Documentation and Conservation* 2(2). 340–350.
- Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel. 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.
- Gordon, Matthew K. 2014. Disentangling stress and pitch accent: Toward a typology of prominence at different prosodic levels. In Harry van der Hulst (ed.), *Word Stress: Theoretical and Typological Issues*. Cambridge: Cambridge University Press.
- Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge 2nd edn.
- Himmelmann, Nicholas P. 2004. Documentary and descriptive linguistics. In Osamu Sakiyama & Fubito Endo (eds.), *Lectures on endangered languages 5: from the Tokyo and Kyoto Conferences 2002*, 37–83.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196.
- Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*. 1–30. Berlin: Mouton de Gruyter.
- Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation* 6. 187–207. <http://hdl.handle.net/10125/4503>.
- Levinson, Stephen C. 2006. On the human ‘interactional engine’. In Stephen C Levinson & Nicholas J Enfield (eds.), *Roots of human sociality culture, cognition and interaction*. New York: Berg Publishers.
- Maskikit-Essed, Raechel & Carlos Gussenhoven. 2016. No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay. *Phonology* 33(02). 353–389. doi:10.1017/S0952675716000154.

- McDonnell, Bradley. 2016. *Symmetrical Voice Constructions in Besemah: A Usage-based Approach*. Santa Barbara: University of California, Santa Barbara Dissertation.
- McDonnell, Bradley. forthcoming. Pragmatic Annotation in the Documentation of Conversation. In Rich Sandoval & Nicholas Jay Williams (eds.), *Interactional Approaches to Language Documentation* (Language Documentation & Conservation Special Publication), .
- McDonnell, Bradley & Rory Turnbull. 2018. Neural network modeling of prosodic prominence in Besemah (Malayic, Indonesia). *Speech Prosody 2018* doi:10.21437/SpeechProsody.2018-154.
- Rhodes, Richard A., Lenore A. Grenoble & Anna Berge. 2006. Adequacy of documentation: A report to the CELP. Ms.
- Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 213–252. Berlin, New York: Mouton de Gruyter. doi:10.1515/9783110197730.213.
- Selting, Margret & Elizabeth Couper-Kuhlen (eds.). 2001. *Introducing interactional linguistics* (Studies in Discourse and Grammar 10). Amsterdam: John Benjamins.
- Sugita, Yuko. 2011. Bringing ‘interactivity’ into language documentation studies. In Peter K. Austin, Oliver Bond, Lutz Marten & David Nathan (eds.), *Language Documentation & Linguistic Theory*, vol. 3, 267–277. London: SOAS.
- van Heuven, Vincent J., Lilie Roosman & Ellen van Zanten. 2008. Betawi Malay word prosody. *Lingua* 118. 1271–1287.
- van Zanten, E. & Robertus Wilhelmus Nicolaas Goedemans. 2007. A functional typology of Austronesian and Papuan stress systems. In Ellen van Zanten & Vincent J van Heuven (eds.), *Prosody in Indonesian Languages*, vol. 9. 63–88. LOT.
- van Zanten, Ellen & Vincent J. van Heuven. 1998. Word stress in Indonesian: Its communicative relevance. *Bijdragen tot de Taal-, Land-en Volkenkunde* 154(1). 129–144.
- van Zanten, Ellen & Vincent J. van Heuven. 2004. Word stress in Indonesian: Fixed or free. *NUSA, Linguistic studies of Indonesian and other languages in Indonesia* 53. 1–20.
- Woodbury, Anthony C. 2011. Language Documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 159–176. Cambridge: Cambridge University Press.
- Yu, Kristine M. 2014. The experimental state of mind in elicitation: illustrations from tonal fieldwork. *Language Documentation & Conservation* 8. 738–777. <http://hdl.handle.net/10125/24623>.

Bradley McDonnell
mcdonn@hawaii.edu

 orcid.org/0000-0001-6422-2022

**Views on Language
Documentation from Around
the World**

Reflections on linguistic fieldwork

Claire Bower
Yale University

In this reflections piece, I draw upon my experience as a fieldworker in Australia, a linguist who also works with archival materials spanning 150 years, and a linguist whose work includes both documentary and descriptive aspects. I center this piece around three questions about aspects of fieldwork that have changed since the publication of Himmelmann (1998). The first is what we collect – that is, have our field methods changed? The second question concerns the documentation we produce – is it different? Thirdly, are there features of Himmelmann’s manifesto which were the products of its time, and has academia changed? Arguably in all cases that there has been change for the better, but we still have some way to go, and that some of the original formulation of a dichotomy between documentation and description are counterproductive.

1. Introduction¹ In this reflections piece, I draw upon my experience as a fieldworker in Australia, a linguist who also works with archival materials spanning 150 years, and a linguist whose work includes both documentary and descriptive aspects. I began fieldwork in 1999, and so my entire professional career as a fieldworker has been in what we might call the “post-Himmelmann” era of language documentation. That is, it has been conducted in the intellectual environment of explicit discussions of field methodology, of documentation as a practice distinction from linguistic analysis, and with in-depth discussion of what it means to work ethically with communities, speakers, and language data.

I center this piece around three questions about aspects of fieldwork that have changed since the publication of Himmelmann (1998). The first is what we collect – that is, have our field methods changed? I argue that field methods have, indeed, changed in several ways. There is more interdisciplinary work; more collaboration with language communities, and more recognition of what needs to go into a documentation project for communities. However, there is still a lot more to say about what we describe when linguists work “on a language”. Himmelmann (1998:161-63, 166) defines the basic object

¹Many thanks to two anonymous reviewers whose comments helped me refine discussion of several topics presented here.

of description as being “observable *linguistic behavior*, manifest in everyday interaction between members of the speech community, and 2) native speakers’ *metalinguistic knowledge*, manifest in their ability to provide interpretations and systematizations for linguistic units and events.” I argue that this does not fit all cases and we should not confine ourselves to field situations where both these objects of description are accessible.

The second question concerns the documentation we produce – is it different? Do we see more corpora (primary descriptive materials) being produced and published? Have we changed what we are doing with the primary resources? I argue that much greater availability of digital recording devices has been the driver of change here, along with a greater focus on archiving as distinct from publishing, but changes are slow.

Thirdly, are there features of Himmelmann’s manifesto which were the products of its time, and has academia changed? Again, there has been change: the dichotomy between ‘community’ and ‘linguist’ is not quite as pronounced as it was 20 years ago; there are more linguists from endangered language communities working on their own languages, and the partnership that Wilkins (1992) discusses for Australia is a more usual way to conduct research, in both the US and Australia and other parts of the world. There is less work that takes no account of community dynamics and pressures. But we could, and should, be doing better. These points have relevance to the objects of documentation, as well as what is done with documentary materials.

2. Have field methods changed? I see several important changes in the type of fieldwork typically undertaken by linguists over the past twenty years. The first is the increasing use of semi-structured elicitation tasks in basic language documentation. Perhaps most famous are the Max Planck Institute (Nijmegen) field manual kits released over the period 1992-2010.² The kits provide visual stimuli, questionnaires, and structured, consistent tasks for use with speakers. Although originally developed for MPI-internal comparative/typological research projects, both the specific stimuli and the general approach have been used as a way of getting controlled data without the prompting of translation-based sentences or the need for speakers to be fluent in a contact language. Further discussion of these methods can be found in Bochnak & Matthewson (2015) and Cover & Tonhauser (2015) amongst others.

The second difference has been the emphasis on the collection of conversational and natural data as part of a documentation project, even when the main goal of the language documentation project is not conversation analysis. While linguists have long made use of a variety of methods for gaining information about the language (not least, participant observation and learning the language; cf. Hale (2001) “do whatever it takes to learn the language”), linguists are now both more explicit about their documentation methods and are using more approaches consistently and deliberately. Indeed, so concerned is Thieberger (2012) about the type of material that is missing from traditional guides that only one chapter in that handbook (Mosel 2012’s ‘guide to the guides’) covers what might be called the core of “traditional” fieldwork. While Abbi (2001), Bower (2008), Crowley (2007), and others are all different books, they do cover much of the same general material, with a focus on language documentation through subfields of linguistics.

Impressionistically, there has been a methodological “smoothing” over the last twenty years. As a graduate student, I remember heated debates about whether descriptions

²These handbooks were originally distributed in print with CDs. They are now available from <http://fieldmanuals.mpi.nl>.

should be based on elicited data or conversational data (alone); each camp firmly convinced of the uselessness of the type of data produced by the other's methods. In brief, elicited data was considered too contaminated by the meta-language of elicitation to be a "true" reflection of data from the language, while conversational data was considered too unstructured and incomplete to be useful at elucidating the internal grammatical competence of an individual. We have a better and more nuanced understanding of the advantages and disadvantages of different types of data – that elicitation and translation provides better evidence of the possible and impossible structures of languages, while natural interactions provide a valuable source of structural data, as well, of course, data about the linguistic behavior itself (more on that below). No doubt, this has come about through more outlets for explicit discussion of field methodology, such as through the pages of journals such as *Language Documentation & Conservation* and *Language Documentation and Description*, as well as recent handbooks (Austin & Sallabank 2011; Thieberger 2012; Chelliah & de Reuse 2010).

In Australia, there are more genuinely interdisciplinary projects. Evans (2012) provides an overview, and Thieberger (2012) has numerous chapters about linguistic knowledge of cultural practices ranging across the natural world. An example of one from Australia is Glenn Wightmann's ethnobiological collaborations with communities across the Top End of Australia (for example, 1994; Wightman, Roberts & Williams 1992). It must be noted, though, that interdisciplinary work isn't new in Australia. Some of the earliest intensive academic fieldwork, such as the Cambridge Expedition to the Torres Strait of 1898, was "interdisciplinary" in that they included scholars from many fields who worked together.

Perhaps the biggest difference between field methods from twenty years ago and today is the use of digital recording technology. Being able to record digitally has made a huge difference to the amount of material that can be recorded, to the workflow for processing recordings, and to the type of work that can be done with such material. The limitations of analog media for recording included the expense of the tapes, the amount of space they took up in luggage to and from the field, the difficulty of making backups (usually they had to be backed up in real time, unless one had a tape deck which could play and record at faster than 1x speed), the cumbersomeness of using them for transcription (play a sentence, rewind, play again), and the fragility of the media (don't leave them in the car on a hot day, or near a magnet). Let alone video. As a result, linguists now record a lot more of their sessions, and the audiovisual component of a documentation project is the richer for it. However, extensive audiovisual recordings bring the problems that come with multiple gigabytes of data, the bottleneck of transcription, and material of incidental relevance to the project. But these problems are small compared to the advantages of being able to time-align recordings with transcriptions, being able to search over multiple corpora and sessions within a second, and being able to extract and manipulate data for phonetic analysis.

In summary, we have changed what we collect, but not, I think, because (or purely because) of the language documentation/description divide which Himmelmann (1998) focused on. Rather, tools like the MPI field manuals have brought to greater prominence the role of semi-structured data gathering (Bower 2008); we have better procedures for interdisciplinary data gathering, and digital data collection and processing tools have reshaped workflows. (For a view that ties this shift to Himmelmann more closely, see Good 2011.) Semi-structured data, however, is seldom useful for maintenance of culture, or linguistically mediated cultural practices.

Before moving to the second question, we should consider the focus of documentation. Himmelmann's focus is on documenting linguistic behavior as the way to document the language. That is, he makes an explicit difference between the recording of linguistic forms which provide evidence for language structures, and the behavior of individuals when talking. I associate this focus particularly with linguists trained within the MPI Nijmegen tradition, and the approach is well illustrated by the field manuals mentioned above. However, this dichotomy is problematic for some field situations. For example, I typically work with speakers who do not use the language we are documenting every day. We have discussions about translations, speakers produce words, sentences, and connected speech in the language, but we document language and linguistic forms with no assumption that what we are doing transfers to a more general set of speech practices. Those speech practices are gone; they've been replaced by a very different linguistic ecology. The majority of the time that the language is used is in these linguistic sessions—that is, the documentation *is* the primary linguistic ecology for these languages these days.

3. Has the production of documentation changed? While our data storage methods have changed greatly with digital recorders, and some types of linguistic data are much easier to collect than they used to be, my impression is that the core of documentation—at least for the linguistic community—is still based around categories of traditional grammar and functional typology. We are getting better data, from more varied sources; we are paying more attention to variation. Handbooks and publications like Thieberger (2012) and Bochnak & Matthewson (2015) have made us more aware of the methods we employ in field work and how to collect data better. But our publications are little changed from the grammars, dictionaries, and texts of 50 years ago. Even the digital grammar collections, like Pacific Linguistics' online *Asia Pacific Linguistics* documentation series, are essentially print books online. We could be doing a lot more to take advantage of the possibilities that digital media allows, such as audiovisual and text linking, or non-linear presentation.

Moreover, since Wilkins' (1992) article on linguistic research under Aboriginal control, a series of papers on ethics, community involvement, and language pedagogy have made documentary linguists more aware of issues of audience, of the difference between pedagogical materials and descriptive grammars, of the ways in which linguistic terminology produces barriers to understanding, and how we can partner with education specialists to improve the materials we produce for and with communities (cf. Czaykowska-Higgins 2009). Simply put, a "community contribution" in the context of an endangered language documentation project is no longer satisfied by a copy of a \$300 reference grammar in the local library. While we are still very focused on "the community" (without appropriate recognition that communities are groups of individuals, not all of whom may agree with one another), we are doing a better job at recognizing what an appropriate and meaningful contribution to a community might look like, and that such contributions will differ depending on the field site (see further Dobrin & Schwartz 2016). Now that the first generation of linguists trained under Wilkins' (1992) model are now senior members of the field, we can see the ways in which (for Australia at least) we have lasting recognition that community linguists and language activists are the crucial drivers of language projects.

Another big difference from 20 years ago is the greater emphasis on archiving, the ethics and responsibilities that linguists have to archive their data; the distinction between archiving and publication (even publication on the web, which was still rare in 1998).

Australia was ahead of the game here, so we see less difference over the last twenty years than in other places. Australia already had an excellent (print) archive, the library of the Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS). Other historical collections are held in regional Museums and the National Library of Australia, but the AIATSIS archive is unique in having an extensive and continuing collection of both print and audiovisual resources for Australian languages. Australia also had ASEDA (Aboriginal Studies Electronic Data Archive), which unfortunately no longer exists (much of its holdings has been transferred to AILEC, housed by AIATSIS). Unfortunately, Australia's digital archiving is perhaps now somewhat behind, compared to AILLA and the Berkeley Survey of California and other Indian Languages. Recent digital field archives for Australian documentation projects are spread between ELAR, Paradisec, and DoBeS (sometimes in several of these archives simultaneously). I suspect a great deal of digital documentation material (particularly secondary analytical materials such as conference handouts, slides, or posters, and data from language revitalization events and language classes) is not being archived.³

4. Has academia changed? Himmelmann (1998) places the linguist at the center of the documentation process. One difference from 20 years ago is the blurring of lines between data and analyses generated by the linguist and those produced by native speakers. A great deal more language documentation is conducted either by linguists who are also native speakers, or with non-native speaker linguists in conjunction with speakers. However, as Hill (2002) and Davis (2017) have shown, linguists talking about endangered languages still do so in a way that constructs barriers to community members. As linguists continue to discuss ways in which our field could improve, a greater commitment to diversity and inclusion (including the ways in which linguists' practices end up excluding the very people they are aiming to hire) should feature.

Students get much better training in field methods, including archiving, ethics, data, etc. When I started fieldwork, the only textbooks were from the 1960s, apart from Vaux & Cooper (1999), which was problematic for work that was community-based. We now have a wealth of material about what fieldwork is, how to do it, what the ethical implications are, and how to deal with data throughout the documentation process.

5. Further reflections and conclusions A few other points are warranted. Perhaps most important is Himmelmann's definition of language documentation (Himmelmann 1998:166). As briefly discussed above, Himmelmann distinguishes between the recording of linguistic behavior and the recording of linguistic judgments. This is problematic. First, it overlooks the fact that for undocumented or underdocumented languages where the linguist has limited field time, it may be most expedient to structure the language documentation around the analytical results (or the description of another, closely related language) or around the compilation of a dictionary. In fact, several fieldwork books advocate working this way, at the same time as endorsing the approach of a separation of language documentation and description. A wordlist with example sentences is a good way to get enough preliminary data for a sketch (compare Hale 2001). I suspect that linguists, on the whole, plan their documentary activities—at least initially—around the end products. Moreover, data collected without an aim (or hypothesis) can be at

³<https://zenodo.org/communities/australianlanguages/> is a free community portal for Australian languages where material of this type can be uploaded, particularly for the archiving of "grey" literature.

best problematic, or at worst, useless. It might be fine for some purposes, but specific hypotheses usually require specific types of data. These goals should be consistent with community expectations of the documentation project as well. For example, if a linguist is brought in under the assumption that the project will document language in use, they should not simply make a wordlist.

“As long as collection and analysis are considered part of a single, uniform, project, the collection activity is likely to be (relatively) neglected” (Himmelmann 1998:164). More primary data are available, thanks to online archives. However, it needs to be acknowledged that primary data are still difficult to use without familiarity with the language (or one closely related). For example, as beautifully laid out and user-friendly as the online Ainu⁴ corpus is, realistically, for most research purposes, I will be looking for an Ainu grammar. And because there are so many languages and so few linguists working on them, it’s often the case that the only person with the requisite knowledge to use raw data from a corpus collection is the linguist who collected it in the first place. So, linguists end up using the secondary sources anyway, even if the raw data are available. Another example is the Chirila database of Australian wordlists (Bowerm 2016); because the scope of the material is Australia-wide, we could not realistically work from audio or unprocessed field notes, but have had to initially prioritize analyzed or at least partially processed sources. There are, however, other use cases where availability of raw materials is preferable. Communities using and adapting materials, for example, are likely to need the primary data.

How much of the changes discussed here are due to the influence of Himmelmann (1998) alone? It’s hard to say. Certainly, the paper has been influential, highly cited, much discussed in the literature on language documentation, and has been accompanied by other very influential publications on fieldwork and recording language. Yet it came at the right time for other changes in academia too – particularly the influence of digital resources on documentary methods. Benchmark papers like this let us see how far we’ve come, but they should not either prevent us from seeing what came before, or stop us from re-imagining 21st Century fieldwork as documentation and description that works with and enhances communities even further.

⁴<http://ainucorpus.ninjal.ac.jp/corpus/en/>


References

- Abbi, Anvita. 2001. *A manual of linguistic field work and structures of Indian languages*. Munich: Lincom Europa.
- Austin, Peter K. & Julia Sallabank (eds.). 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.
- Bochnak, Ryan & Lisa Matthewson. 2015. *Methodologies in Semantic Fieldwork*. Oxford University Press.
- Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. Basingstoke: Palgrave Macmillan.
- Bowern, Claire. 2016. Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia. *Language Documentation & Conservation* 10. 1-44.
- Chelliah, Shobhana & Willem de Reuse. 2010. *Handbook of Descriptive Linguistic Fieldwork*. New York: Springer.
- Cover, Rebecca & Judith Tonhauser. 2015. Theories of meaning in the field: Temporal and aspectual reference. In Ryan Bochnak & Lisa Matthewson (eds.), *Methodologies in Semantic Fieldwork*, 306–349. Oxford University Press.
- Crowley, Terry. 2007. *Field linguistics: A beginner's guide*, ed. by Nick Thieberger. Oxford University Press.
- Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation & Conservation* 3(1). 15-50. <http://hdl.handle.net/10125/4423>
- Davis, Jenny L. 2017. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. *Language Documentation and Description* 14. 37-58.
- Good, Jeff. 2011. Data and language documentation. Cambridge University Press. In Peter Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 212–234. Cambridge: Cambridge University Press.
- Hale, Kenneth L. 2001. Ulwa (Southern Sumu): The beginnings of a language research project. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 76–101. Cambridge: Cambridge University Press.
- Hill, Jane H. 2002. “Expert Rhetorics” in Advocacy for Endangered Languages: Who Is Listening, and What Do They Hear? *Journal of Linguistic Anthropology* 12(2). 119–133.
- Himmelman, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196.
- Thieberger, Nicholas (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford University Press.
- Vaux, Bert & Justin Cooper. 1999. *Introduction to linguistic field methods*. Munich: Lincom Europa.
- Wightman, Glenn M. 1994. *Gurindji ethnobotany: Aboriginal plant Use from Daguragu, Northern Australia*. Darwin: Conservation Commission of the Northern Territory.
- Wightman, G., J. G. Roberts, & L. Williams. 1992. Mangarrayi ethnobotany aboriginal plant use from the Elsey area Northern Australia. *Northern Territories Botanical Bulletin* 15.

Wilkins, David. 1992. Linguistic research under Aboriginal control: A personal account of fieldwork in Central Australia. *Australian Journal of Linguistics* 12(1). 171–200.

Claire Bowern

claire.bowern@yale.edu

 orcid.org/0000-0002-9512-4393

The state of documentation of Kalahari Basin languages

Tom Güldemann

Humboldt University Berlin

Max Planck Institute for the Science of Human History Jena

The Kalahari Basin is a linguistic macro-area in the south of the African continent. It has been in a protracted process of disintegration that started with the arrival of Bantu peoples from the north and accelerated dramatically with the European colonization emanating from the southwest. Before these major changes, the area hosted, and still hosts, three independent linguistic lineages, Tuu, Kx'a, and Khoe-Kwadi, that were traditionally subsumed under the spurious linguistic concept “Khoisan” but are better viewed as forming a “Sprachbund”. The languages have been known for their quirky and complex sound systems, notably involving click phonemes, but they also display many other rare linguistic features—a profile that until recently was documented and described very insufficiently. At the same time, spoken predominantly by relatively small and socially marginalized forager groups, known under the term “San”, most languages are today, if not on the verge of extinction, at least latently endangered. This contribution gives an overview of their current state of documentation, which has improved considerably within the last 20 years.

1. The picture 20 years ago¹ The languages under discussion, formerly known simply as “click” languages, had been commonly subsumed until recently under the so-called “Khoisan” family, following Greenberg (1963). Today specialists no longer follow this premature genealogical proposal and increasingly work with an areal hypothesis called

¹I am grateful to Hiroshi Nakagawa and Bonny Sands for furnishing information on relevant research in Japan and the USA, respectively. The abbreviations below are: AF Arcadia Fund; DASTI Danish Agency for Science, Technology and Innovation; DFG Deutsche Forschungsgemeinschaft; ESF European Science Foundation; FFAF Firebird Foundation for Anthropological Research; GMF Guggenheim Memorial Foundation; JSPS Japan Society for the Promotion of Science; NSF National Science Foundation; NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek; Ph.D. doctoral dissertation; VWF Volkswagen Foundation.

“Kalahari Basin” (see Güldemann (2014) and Güldemann and Fehn (2017) for the most recent discussion on language classification and areal linguistics, respectively).

Shortly after the appearance of Himmelmann (1998), a survey of the languages by Güldemann and Vossen (2000) drew an alarming picture, this not only about the precarious sociolinguistic situation but also the deficient state of documentation of the languages in question. The information given in that article is repeated in Table 1, disregarding both extinct languages and languages spoken in eastern Africa outside the Kalahari Basin. It shows that at that time only three out of about 20 languages were sufficiently documented and described by means of publicly available material, namely Kxoe aka Caprivi Khwe, Nama-Damara aka (mainstream) Namibian Khoekhoe, and the Jul’hoan dialect of Ju.

No.	Language	Phonetics/ phonology	Lexicon	Grammar	Raw texts	Glossed texts
3	Hiecho		(S)	(S)	(S)	
5	Kxoe		S U	M	M	M
	Buga, !Ani	S	S U	S	U	U
6	G!ui, G!lana	S	S			
7	Naro	S	M M	S		
9	!Ora	M	(S)	(M S)	(M S)	
11	Nama-Damara	M S	U (M S)	M T (M)	U (M)	
12/3	Hailom-†Aakhoe		U	U	U	(S)
14	!Xūū	S	S	M	(S U)	S U
	Jul’hoan	M S	M M	M M M	S U	
15	†Hōā		S	S		
16	!Xōō	M S	M	S (U)		U

Table 1: Documentation state of major languages of the Kalahari Basin area around 2000 (after Güldemann and Vossen 2000: 103). **Note:** No. = language key to Table 2; M = monograph; S = short treatment; T = thesis; U = unpublished manuscript; (...) = outdated; Shading = good description

Since the time of this publication the situation has changed considerably. While this can unfortunately not be said concerning language vitality, it certainly holds for the state of documentation, as discussed in the following.

2. The task One reflex of the problems existing 20 years ago is that Güldemann and Vossen (2000: 99) still operated with the largely unclear issue of language classification and, in speaking of “thirty or so” living “languages and dialects”, with an indeterminate language inventory. Today, the overall situation enjoys more clarity, as discussed by Güldemann (2014). For one thing, Greenberg’s idea of a single language family has been widely abandoned. Moreover, the number of relevant languages and language complexes that are attested or can be assumed to have still been spoken in the Kalahari Basin in the second half of the last century can be established at around 20, as given in the updated list of Table 2 and shown in Figure 1.

Unfortunately, a number of languages are by now extinct or at least moribund. These are Kwadi; !Ora-Xiri and Eini of Khoekhoe (all of Khoe-Kwadi); †Amkoe (of Kx’a); as well

as N!ng, !Xegwi and the Lower Nossob group (all of Tuu), although †Amkoe and N!ng are still subject to fieldwork. This reduces the languages that can be analyzed today with the help of native speakers to about 15, as to be discussed in the following (see Table 3 for a full list). The different inventory compared to other studies, for example, Brenzinger (2013) with 10 languages, is mostly due to persisting problems concerning the notorious language-dialect distinction, particularly in the Khoe family (cf. Güldemann 2014: 6-9).

Family	No.	Language (complex)	Language name in Ethnologue	ISO
Khoe-Kwadi	1	Kwadi [°]	Kwadi	kwz
	2	<i>Shua</i>	Shua	shg
	3	<i>Tshwa</i>	Kua	tyu
			Tsoa	hio
	4	Ts'ixa	under !Ani	–
	5	<i>Khwe</i>	Khwe	xuu
			!Ani	hnh
	6	<i>G!ana</i>	!Gana	gnk
			!Gui	gwj
	7	<i>Naro</i>	Naro	nhr
	8	<i>!Ora-Xiri</i> [°]	Korana	kqz
			Xiri	xii
	9	<i>Eini</i> †	–	–
	10	<i>Nama-Damara</i> *	Khoekhoe	naq
	11	Hai!lom*	Hai!lom	hgm
	12	†Aakhoe*	under Hai!lom	–
Kx'a	13	<i>Ju</i>	Jul'hoan	ktz
			Kung-Ekoka	knw
			Northwestern !Kung	vaj
	14	†Amkoe [°]	†Hua	huc
Tuu	15	<i>Taa</i>	!Xóõ	nmn
	16	!Auni†	–	–
	17	!Haasi†	–	–
	18	<i>N!ng</i> [°]	N!u	ngn
	19	!Xegwi†	!Xegwi	xeg

Table 2: Languages and language complexes of the Kalahari Basin area spoken in the second half of the 20th century (after Güldemann 2014). **Note:** *italic* = language complex/dialect cluster, † = extinct, ° = moribund, * = subsumed under Standard Namibian Khoekhoe aka 'Khoekhoegowab'

3. The advances Compared to the detrimental state of research around 20 years ago represented in Table 1, the situation has improved immensely, which is due to various factors. For one thing, since the late 1990s the community involved in so-called “Khoisan” research has been meeting at a regular conference series, initiated and organized for many years by Rainer Voßen and Bernd Heine, notably in 1997, 2003, 2006, 2008, 2011, 2014,

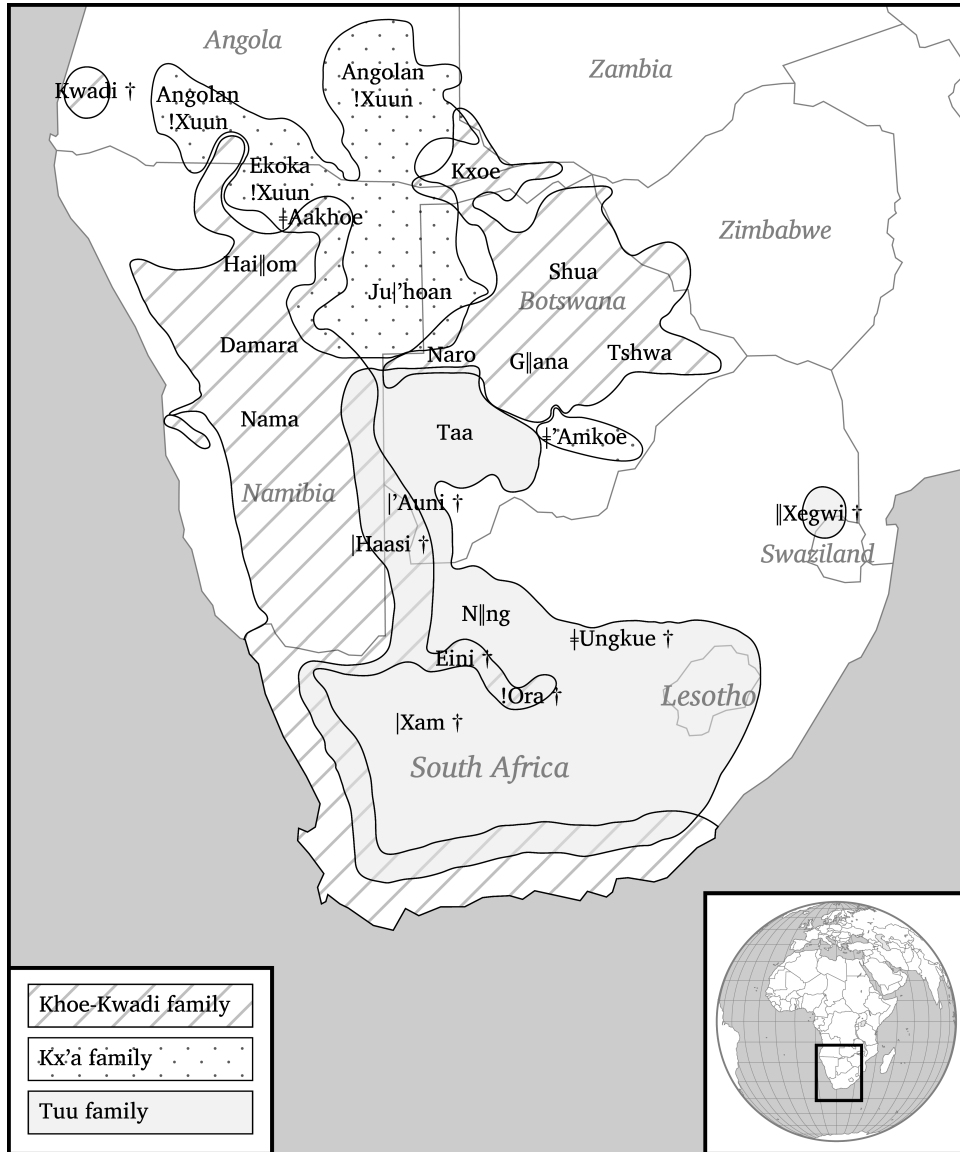


Figure 1: Map of the approximate distribution of the Kalahari Basin languages.

2017. This has intensified academic exchange and helped to define and coordinate research agendas, among them language documentation.

In terms of publications, these events have until now resulted in six volumes of proceedings, Schladt (1998), Ermisch (2008), Brenzinger and König (2010), Witzlack-Makarevich and Ernszt (2013), Shah and Brenzinger (2016), and Fehn (2017). These and a number of monographs have contributed to the fact that the previously established book series “Research in Khoisan Studies,” published today by Rüdiger Köppe Verlag, has increased by 20 new volumes since 2000.

Another major publication achievement is the long-awaited “Khoesan” handbook edited by Vossen (2013). Although the volume appeared with a great delay, resulting in its contents not reflecting the current state of research at its publication date, and it has a series of other shortcomings (see McGregor 2016a), it provides for the first time a comprehensive overview of all relevant languages except Hailom and N!ng (cf. Table 3).

Intensified research, including that from southern African scholars, increased the knowledge about the sociolinguistic and demographic status of the languages, which also contributed to a better understanding of their dialectological complexity (cf., e.g., Hasselbring 2000; Hasselbring, Segathle and Munch 2000; Crawhall 2004; Haacke 2005; Nakagawa 2006b; Rapold & Widlok 2008; Killian 2009; Gerlach and Berthold 2011; Brenzinger 2013; Naumann 2014; Güldemann 2017).

The major impetus for the greatly intensified documentation activities was given without doubt by the growing recognition of language endangerment on a global scale and specifically in the Kalahari Basin area. This resulted in various initiatives by larger funding bodies to provide financial opportunities for the scientific rescuing of some amount of this dwindling linguistic diversity. This will be discussed in the rest of this section.

Two major projects on †Aakhoe (Khoe-Kwadi) and Taa (Tuu) were carried out within the German VWF-funded program “Documentation of Endangered Languages (DOBES)” (see <http://dobes.mpi.nl/>). The “Endangered Languages Documentation Programme (ELDP)” in London (see <http://www.eldp.net/en/our+projects/projects+list/>) has so far funded or still funds major research projects on N!ng (Tuu), Jul’hoan (Kx’a), and Tshwa (Khoe-Kwadi) as well as minor projects on Mangetti Dune !Xung (Kx’a) and !Ora (Khoe-Kwadi). Four languages, Taa (Tuu), †Amkoe (Kx’a), Shua, and !Ora (both Khoe-Kwadi), received major attention within the “Kalahari Basin area (KBA)” project (see <https://www2.hu-berlin.de/kba/projects.html>) as part of the program “Better Analyses Based on Endangered Languages (EuroBABEL)” funded by the ESF and four national agencies. US American funding bodies like the NSF, GMF, and FFAF supported research on Jul’hoan (Kx’a) and N!ng (Tuu) as well as Khoisan phonetics encompassing various languages. Finally, the JSPS in Japan has been funding detailed documentation on G!ana and Naro (both Khoe-Kwadi) as well as research on Khoisan phonetics.

All this intensive engagement has resulted in around ten Ph.D. theses (Crawhall 2004, Nakagawa 2006, Exter 2008, Brugman 2009, Killian 2009, Mathes 2015, Fehn 2016, Gerlach 2016, Pratchett 2017) some of whose authors are rejuvenating the research community.

Another positive development is that anthropological research, which traditionally has been very active in the area yet not very attentive to language-related topics, has now also started to contribute more to such documentation. This is evident from studies like Boden and Michels (2000) and Boden (2001) on Caprivi Khwe (Khoe-Kwadi), Tanaka and Sugawara (2010) on G!ana-G!ui (Khoe-Kwadi), and Barnard and Boden (2014) on kinship systems of the entire area.

All the progress mentioned above is summarized in Table 3, which presents the current documentation status of Kalahari Basin languages and can be directly compared with Table 1, which reflects the situation 20 years ago. Table 3 records the recent research projects dedicated to individual languages as well as the publications and archival or database deposits that have become available through them. The latter material is separated according to Himmelmann's (1998) trilogy of lexicon, grammar, and (raw vs. linguistically annotated) texts but additionally singles out phonetics-phonology, involving in particular experimental phonetics. This is because the languages are so complex in this last area that they cannot be viewed as fully analyzed without such a dedicated treatment. This becomes evident by the fact that the first appropriate analyses of solely the phonetics-phonology of such languages as Ju'hoan, East !Xoon, and G!ui by Snyman (1975), Traill (1985), and Nakagawa (2006), respectively, involved research periods of ten years and more. The complexity does not only concern the typologically quirky clicks but also other rare consonants and suprasegmental vowel features, leading to some of the most complex phoneme systems on a global scale (see Güldemann and Nakagawa (2018) for a recent discussion on some typological issues).

If a subdomain is treated for an individual language by one or more published monographs or an accessible database, it can be considered to be well documented and described (marked by shading in Table 3). This situation is normally accompanied by the availability of additional detailed articles on special topics, which is not exhaustively reflected in the table. Where a monograph on grammatical or phonetic-phonological description is not yet available, I give sample articles; this list is normally not exhaustive but only shows that the language is in the process of being analyzed. It should also be recognized that the equally growing published outcome of comparative research also contains a good amount of language-specific data not yet accessible in larger language-specific studies.

In general, Table 3 clearly demonstrates the major progress compared to a mere three languages that were reasonably described 20 years ago. Of the 14 relevant languages and language complexes, six are by now well known, namely Khwe and Nama-Damara of Khoe-Kwadi, Ju and moribund †Amkoe of Kx'a, and Taa and moribund N!ng of Tuu. Another six, namely Shua, Tshwa, Ts'ixa, G!ana, Naro, and †Aakhoe, all members of Khoe-Kwadi, range from extensively to at least reasonably well documented by recent research, although the results are not yet fully published and/or archived, thus remaining publicly inaccessible. The only language units where modern scholarship is very deficient still today are Hailom and !Ora-Xiri (both from Khoe-Kwadi); for the latter, this is beyond remedy, as the work with remaining (semi)speakers started too late.

4. The future Against the background of the present state of documentation and description sketched in §3, a few points can be made regarding the future work that is ahead of the specialists studying the languages of the Kalahari Basin area.

In terms of basic language coverage, a somewhat unexpected result is that the Khoekhoe variety of the Hailom, a generally well-known group of earlier foragers around the Etosha Pan (see Friederich 2009), remains all but unknown, whereby its linguistic status as a dialect or language is still unclear (cf. Haacke, Eiseb and Namaseb 1997). It may be confusing in this respect that linguistic publications referring to this language name do exist; in fact, they are on the yet different †Aakhoe variety of Khoekhoe, which has been researched intensively by Terttu Heikkinen and subsequently by a major DOBES project.

1	2	3	4	5	6	7	
No.	Language (complex)	Documentation Project	Phonetics/Phonology	Lexicon	Grammar	Raw texts	Glossed texts
2	<i>Shua</i>	ESF DASTI			McGregor (2014, 2015, 2017)		
3	<i>Tshwa</i>	Ph.D., AF, GMF	Snyman (2000), Chebanne (2000, 2013), MATHES (2015)		Chebanne (2008, 2013), Chebanne & Collins (2017), Fehn & Phiri (2017)	MAD	MAD
4	Ts'ixa	Ph.D.			FEHN (2016)		
5	<i>Khwe</i>	DFG		KILLIAN-HATZ (2003)	M. KILLIAN-HATZ (2008)	M	M. HEINE (2010), BODEN (2014)
6	<i>Glana</i>	Ph.D., JSPS	NAKAGAWA (2006a)		Ono (2010), Nakagawa (2013, 2016)		
7	<i>Naro</i>	JSPS			Haacke (2010), Visser (2010)		
8	<i>'Ora-Xiri</i> *	ESF NWO, AF	Nakagawa (2017)	M, VISSER (2001)		M	
10	<i>Nama-</i>	Ph.D.	M, BRUGMAN (2009)	HAACKE & EISEB (2002)		M	
11	<i>Damara</i> *						
12	<i>!Aakhoë</i> *	VWF					
13	<i>Ju</i>	Ph.D., AF, NSF, FFAR	M, MILLER-OCKHUIZEN (2003)	M. KÖNIG & HEINE (2008)	Widlok, Rapold & Hoymann (2008), Widlok (2008, 2016), Hoymann (2010), Rapold (2012), Haacke (2013)	MAD, Schmidt (2011)	MAD
14	<i>f'Amkoe</i> ^o	Ph.D., ESF DFG, GMF			M. DICKENS (2005), KÖNIG & HEINE (2001), HEINE & KÖNIG (2015), PRATCHETT (2017)	BIESELE (2009), Schmidt (2011)	MAD
15	<i>Taa</i>	VWF, ESF DFG	GERLACH (2016) M. Naumann (2008, 2017)		COLLINS & GRUBER (2014)	MAD	MAD
18	<i>Ning</i> ^o	NSF, AF	Miller et al. (2007), EXTER (2008)	M. MAD, TRAILL (2018) MAD	Kiefling (2008, 2013, 2017) COLLINS & NAMASEB (2011), Ernszt, Witzlack-Makarevich & Güldemann (2015)	MAD	MAD

Table 3: Documentation state of spoken languages and language complexes of the Kalahari Basin area. **Note:** No. = language key to Table 2; *italic* = language complex; **bold** in column 1 = not treated in Vossen (2013); **bold** in column 2 = more than 1 project; Columns 3-7: CAPITALS = MONOGRAPH, M = monograph(s) before 2000, MAD = modern archival deposit, Shading = good and publicly available documentation/description, ^o = moribund, * = subsumed under Standard Namibian Khoekhoe.

This open problem of dialectal diversity points to a more general persisting deficit in the field. As pointed out above, among the 14 units of Table 3, there are a number of language complexes in terms of Hockett (1958), some of which display an internal heterogeneity amounting at times to mutual unintelligibility that is far from being understood fully. Thus, a better coverage of dialect diversity is imperative for a conclusive assessment of the language distinctions in the area.

Given the relatively recent and thus still restricted linguistic engagement with the Kalahari Basin languages, it goes without saying that scholarship needs to broaden the range of linguistic topics studied. To give just one example, studies on lexical semantics or on language and cognition are still limited in the field (for a few exceptions, see Brenzinger (2008), Widlok (2008) and McGregor (2016b) on spatial language; Nakagawa (2012) and Brenzinger and Fehn (2013) on the domain of perception verbs; and McGregor (2014) on numeral conceptualization).

Regarding future tasks concerning the research that has been achieved already, two points come to mind in particular. For one thing, there is a considerable amount of material that was collected in the past but which requires (more complete) archiving, especially data that were not produced in the framework of a major documentation initiative with the necessary infrastructure, including legacy material of scholars no longer active and/or alive. Moreover, we must not be content with collecting data and storing them in archival deposits but continue to analyze and annotate them in depth, so that they can be used effectively once speakers can no longer be consulted, which is imminent for some of the languages.

Finally, for the benefit of both effective academic exchange and practical issues of speech communities, it is necessary to strive for better and, if possible, unified description and representation standards. In particular, this holds a) for similar grammatical phenomena across closely related dialects and languages, which is first of all relevant for the Khoe family (cf., e.g., the discussion revolving around multi-verb constructions), and b) for the complex features of the sound systems that recur across all three lineages of the area (see, e.g., the ongoing controversies revolving around practical orthographies discussed in such works as Güldemann 1998, Snyman 1998, Miller-Ockhuizen 2000, Schladt 2000, Visser 2000, and Namaseb et al. 2008).

Last but not least, one of the central problems in the field is that still too little scholarship comes from researchers from southern Africa itself, and almost none from mother tongue speakers, which is due to their overall low formal education level even for African standards. To support developing local southern African and native speaker scholarship of high quality is thus one of the priorities for our academic community.

References

- Barnard, Alan & Gertrud Boden (eds.). 2014. *Southern African Khoisan kinship systems*. Köln: Rüdiger Köppe.
- Batibo, Herman M. & Joe Tsonope (eds.). 2000. *The state of Khoesan languages in Botswana*. Gaborone: Basarwa Languages Project.
- Biese, Megan (ed.). 2009. *Ju'hoan folktales: Transcriptions and English translations*. Victoria, British Columbia: Trafford.
- Boden, Gertrud. 2001. *Kxoe material culture: Aspects of classification and change with database on CD-ROM* (Khoisan Forum, Working Papers 18). Köln: Universität zu Köln.
- Boden, Gertrud (ed.). 2014. *Khwe kúri-x'ón-dji - Khwe family names*. Köln: Kalahari Cultural Heritage Publications.
- Boden, Gertrud & Stefanie Michels. 2000. *Kxoe material culture: Aspects of change and its documentation, subsistence equipment* (Khoisan Forum, Working Papers 16). Köln: Universität zu Köln.
- Brenzinger, Matthias. 2008. Conceptual strategies of orientation among Khwe: From sunrise/sunset bisections to a left/right opposition. In Sonja Ermisch (ed.), *Khoisan languages and linguistics: Proceedings of the 2nd International Symposium, January 8–12, 2006*, Riezlern/Kleinwalsertal, 15–47. Köln: Rüdiger Köppe.
- Brenzinger, Matthias. 2013. The twelve modern Khoisan languages. In Alena Witzlack-Makarevich & Martina Ernszt (eds.), *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6–10, 2008*, Riezlern/Kleinwalsertal, 1–31. Köln: Rüdiger Köppe.
- Brenzinger, Matthias & Anne-Maria Fehn. 2013. From body to knowledge: Perception and cognition in Khwe-!Ani and Ts'ixa. In Alexandra Y. Aikhenvald & Anne Storch (eds.), *Perception and cognition in language and culture*, 161–191. Leiden: E.J. Brill.
- Brenzinger, Matthias & Christa König (eds.). 2010. *Khoisan languages and linguistics: Proceedings of the 1st International Symposium January 4–8, 2003*, Riezlern/Kleinwalsertal. Köln: Rüdiger Köppe.
- Brugman, Johanna C. 2009. *Segments, tones and distribution in Khoekhoe prosody*. Ithaca: Cornell University dissertation.
- Chebanne, Anderson M. 2000. The phonological system of the Cuaa language. In Herman M. Batibo & Joe Tsonope (eds.), *The state of Khoesan languages in Botswana*, 18–32. Gaborone: Basarwa Languages Project.
- Chebanne, Anderson M. 2008. Person, gender and number markings in Eastern Kalahari Khoe: Existence or traces? In Sonja Ermisch (ed.), *Khoisan languages and linguistics: Proceedings of the 2nd International Symposium, January 8–12, 2006*, Riezlern/Kleinwalsertal, 49–65. Köln: Rüdiger Köppe.
- Chebanne, Anderson M. 2013. Cirecire word morphology and tonology: A preliminary analysis. In Alena Witzlack-Makarevich & Martina Ernszt (eds.), *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6–10, 2008*, Riezlern/Kleinwalsertal, 163–184. Köln: Rüdiger Köppe.
- Chebanne, Anderson M. & Chris Collins. 2017. Tense and aspect in Kua: A preliminary assessment. In Anne-Marie Fehn (ed.), *Khoisan languages and linguistics: Proceedings of the 4th International Symposium July 11–13, 2011*, Riezlern/Kleinwalsertal, 91–108. Köln: Rüdiger Köppe.
- Collins, Chris & Jeffrey S. Gruber. 2014. *A grammar of #Höä*. Köln: Rüdiger Köppe.

- Collins, Chris & Levi Namaseb. 2011. *A grammatical sketch of N|uuki with stories*. Köln: Rüdiger Köppe.
- Crawhall, Nigel. 2004. *!Ui-Taa language shift in Gordonia and Postmasburg Districts, South Africa*. Cape Town: University of Cape Town dissertation.
- Dickens, Patrick J. 2005. *A concise grammar of ǀxǀhoan with a ǀxǀhoan-English glossary and a subject index*. Köln: Rüdiger Köppe.
- Ermisch, Sonja (ed.). 2008. *Khoisan languages and linguistics: Proceedings of the 2nd International Symposium, January 8–12, 2006, Riezlern/Kleinwalsertal*. Köln: Rüdiger Köppe.
- Ernszt, Martina, Alena Witzlack-Makarevich, Tom Güldemann. 2013. N|ng valency patterns. In Iren Hartmann, Martin Haspelmath & Bradley Taylor (eds.), *Valency patterns Leipzig*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online: <http://valpal.info/languages/nllng>)
- Exter, Mats. 2008. *Properties of the anterior and posterior click closures in N|uu*. Köln: Universität zu Köln dissertation. (<http://kups.ub.uni-koeln.de/4979/>) (Accessed 2018-01-05.)
- Fehn, Anne-Maria. 2016. *A grammar of Ts'ixa (Kalahari Khoe)*. Köln: Universität zu Köln dissertation. (<http://kups.ub.uni-koeln.de/7062/>)
- Fehn, Anne-Marie (ed.). 2017. *Khoisan languages and linguistics: Proceedings of the 4th International Symposium, July 11–13, 2011, Riezlern/Kleinwalsertal*. Köln: Rüdiger Köppe.
- Fehn, Anne-Maria & Admire Phiri. 2017. Nominal marking in Northern Tshwa (Kalahari Khoe). *Stellenbosch Papers in Linguistics* 48. 105–122.
- Friederich, Reinhard (ed. Lempp, Horst). 2009. *Verjagt..., verweht..., vergessen...: Die Hailom und das Etoshagebiet*. Windhoek: Macmillan Education Namibia.
- Gerlach, Linda. 2016. *N|aǀqriaxe: The phonology of an endangered language of Botswana*. Wiesbaden: Harrassowitz.
- Gerlach, Linda & Falko Berthold. 2011. The sociolinguistic situation of ǀHoan, a moribund 'Khoisan' language of Botswana. *Afrikanistik online*, vol. 2011. (urn:nbn:de:0009-10-31645)
- Greenberg, Joseph H. 1963. *The languages of Africa*. Bloomington: Indiana University Press.
- Güldemann, Tom. 1998. *San languages for education: A linguistic short survey and proposal on behalf of the Molteno Early Literacy and Language Development (MELLD) Project in Namibia*. Okahandja: National Institute of Educational Development, Ministry of Basic Education and Culture.
- Güldemann, Tom. 2014. "Khoisan" linguistic classification today. In Tom Güldemann & Anne-Maria Fehn (eds.), *Beyond 'Khoisan': Historical relations in the Kalahari Basin*, 1–41. Amsterdam: John Benjamins.
- Güldemann, Tom. 2017. Casting a wider net over N|ng: The older archival resources. *Anthropological Linguistics*. 59(1). 71–104.
- Güldemann, Tom & Anne-Maria Fehn (eds.). 2014. *Beyond 'Khoisan': Historical relations in the Kalahari Basin*. Amsterdam: John Benjamins.
- Güldemann, Tom & Anne-Maria Fehn. 2017. The Kalahari Basin area as a "Sprachbund" before the Bantu expansion. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 500–526. Cambridge: Cambridge University Press.

- Güldemann, Tom & Hiroshi Nakagawa. 2018. Anthony Traill and the holistic approach to Kalahari Basin sound design. In Tom Güldemann & Hiroshi Nakagawa (eds.), *Kalahari Basin sound structure - in memory of Anthony T. Traill (1939-2007)*. *Africana Linguistica* 24. 1–29.
- Güldemann, Tom & Rainer Voßen. 2000. Khoisan. In Bernd Heine & Derek Nurse (eds.), *African languages: An introduction*, 99–122. Cambridge: Cambridge University Press.
- Haacke, Wilfrid H. G. 2005. Linguistic research for literary empowerment of Khoesaaan languages of Namibia. *African Studies* 64(2). 157–176.
- Haacke, Wilfrid H. G. 2010. Naro syntax from the perspective of the desentential hypothesis: The minimal sentence. In Matthias Brenzinger & Christa König (eds.), *Khoisan languages and linguistics: Proceedings of the 1st International Symposium January 4-8, 2003, Riezlern/Kleinwalsertal*, 201–230. Köln: Rüdiger Köppe.
- Haacke, Wilfrid H. G. 2013. On the manifestation of core arguments in †Ákhoe. In Alena Witzlack-Makarevich & Martina Ernszt (eds.), *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6–10, 2008, Riezlern/Kleinwalsertal*, 61–81. Köln: Rüdiger Köppe.
- Haacke, Wilfrid H. G. & Eliphaz Eiseb. 2002. *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*. Windhoek: Gamsberg Macmillan.
- Haacke, Wilfrid H. G. & Eiseb, Eliphaz & Namaseb, Levi. 1997. Internal and external relations of Khoekhoe dialects: A preliminary survey. In Haacke, Wilfrid H.G. & Edward D. Elderkin, (eds.), *Namibian languages: Reports and papers*, 125–209. Köln: Rüdiger Köppe.
- Hasselbring, Sue. 2000. *A sociolinguistic survey of the languages of Botswana*, vol. 1. Mogoditshane: Tasalls.
- Hasselbring, Sue, Thabiso Segathle & Julie Munch. 2000. *A sociolinguistic survey of the languages of Botswana*, vol. 2. Mogoditshane: Tasalls.
- Heine, Bernd. 2010. *Khwe texts*. (Khoisan Forum, Working Papers 8). Köln: Universität zu Köln.
- Heine, Bernd & Christa König. 2015. *The !Xun language: A dialect grammar of Northern Khoisan*. Köln: Rüdiger Köppe.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Hockett, Charles F. 1958. *A course in modern linguistics*. New York: Macmillan.
- Hoymann, Gertie. 2010. Questions and responses in †Ákhoe Hailom. *Journal of Pragmatics* 42. 2726–2740.
- Kießling, Roland. 2008. Noun classification in !Xoon. In Sonja Ermisch (ed.), *Khoisan languages and linguistics: Proceedings of the 2nd International Symposium, January 8–12, 2006, Riezlern/Kleinwalsertal*, 225–248. Köln: Rüdiger Köppe.
- Kießling, Roland. 2013. Verbal serialisation in Taa (Southern Khoisan). In Alena Witzlack-Makarevich & Martina Ernszt (eds.), *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6–10, 2008, Riezlern/Kleinwalsertal*, 33–60. Köln: Rüdiger Köppe.
- Kießling, Roland. 2017. Experiencer encoding in Taa (Southern Khoisan). In Raija Kramer & Roland Kießling (eds.), *Mechthildian approaches to Afrikanistik: Advances in language based research in Africa, Festschrift für Mechthild Reh*, 189–224. Köln: Rüdiger Köppe.
- Killian, Don. 2009. *Khoemana and the Griqua: Identity at the heart of phonological attrition*. Helsinki: University of Helsinki dissertation.

- Kilian-Hatz, Christa. 2003. *Khwe dictionary*. Köln: Rüdiger Köppe.
- Kilian-Hatz, Christa. 2008. *A grammar of modern Khwe (Central Khoisan)*. Köln: Rüdiger Köppe.
- König, Christa & Bernd Heine. 2001. *The !Xun of Ekoka: A demographic and linguistic report* (Khoisan Forum, Working Papers 17). Köln: Universität zu Köln.
- König, Christa & Bernd Heine. 2008. *A concise dictionary of northwestern !Xun*. Köln: Rüdiger Köppe.
- Mathes, Timothy K. 2015. *Consonant-tone interaction in the Khoisan language Tsua*. New York: New York University dissertation.
- McGregor, William B. 2014. Numerals and number words in Shua. *Journal of African Languages and Linguistics* 35(1). 45–90.
- McGregor, William B. 2015. Four counter-presumption constructions in Shua (Khoe-Kwadi, Botswana). *Lingua* 158. 54–75.
- McGregor, William B. 2016a. Review: Vossen, Rainer (ed.). 2013. *The Khoesan languages*. London: Routledge. *Journal of African Languages and Linguistics* 37(1). 165–172.
- McGregor, William B. 2016b. Shua spatial language and cognition: A prolegomenon. In Sheena Shah & Matthias Brenzinger (eds.), *Khoisan languages and linguistics: Proceedings of the 5th International Symposium*, July 13–17, 2014, Riezlern/Kleinwalsertal, 243–276. Köln: Rüdiger Köppe.
- McGregor, William B. 2017. Unusual manner constructions in Shua (Khoe-Kwadi, Botswana). *Linguistics* 55(4). 857–897.
- Miller-Ockhuizen, Amanda L. 2000. Issues in Ju!’hoansi orthography and their implications for the development of orthographies for other Khoesan languages. In Batibo, Herman M. & Tsonope, Joe (eds.), *The state of Khoesan languages in Botswana*, 108–124. Gaborone: Basarwa Languages Project.
- Miller-Ockhuizen, Amanda L. 2003. *The phonetics and phonology of gutturals: A case study from Ju!’hoansi*. New York/ London: Routledge.
- Miller, Amanda L., Johanna Brugman, Bonny Sands, Levi Namaseb, Mats Exter & Chris Collins. 2007. The sounds of N|uu: Place and airstream contrasts. *Working Papers of the Cornell Phonetics Laboratory* 16. 101–160.
- Nakagawa, Hiroshi. 2006a. *Aspects of the phonetic and phonological structure of the G|ui language*. Johannesburg: University of the Witwatersrand dissertation.
- Nakagawa, Hiroshi. 2006b. |Gui dialects and |Gui-speaking communities before the relocation from the CKGR. *Pula* 20(1). 42–52.
- Nakagawa, Hiroshi. 2012. The importance of TASTE verbs in some Khoe languages. *Linguistics* 50(3). 395–420.
- Nakagawa, Hiroshi. 2013. G|ui ideophones: Work in progress. *Asian and African Languages and Linguistics* 8. 99–121.
- Nakagawa, Hiroshi. 2016. The aspect system in G|ui: With special reference to postural features. *African Study Monograph* 52. 119–134.
- Nakagawa, Hiroshi. 2017. †Haba lexical tonology. In Anne-Marie Fehn (ed.), *Khoisan languages and linguistics: Proceedings of the 4th International Symposium*, July 11–13, 2011, Riezlern/Kleinwalsertal, 109–119. Köln: Rüdiger Köppe.
- Namaseb, Levi, Wilfrid Haacke, Laurentius Davids, Blesswell K. Kure, A. Araes, D. Ortman, N. Fredericks, & M.C. Moeti. 2008. *The standard unified orthography for Khoe and San languages of southern Africa* (CASAS Monograph Series 232). Cape Town: Center for Advanced Studies of African Society (CASAS).

- Naumann, Christfried. 2008. High and low tone in Taa †aan (!Xóó). In Ermisch, Sonja (ed.), *Khoisan languages and linguistics: Proceedings of the 2nd International Symposium*, January 8-12, 2006, Riezlern/Kleinwalsertal, 279–302. Köln: Rüdiger Köppe.
- Naumann, Christfried. 2014. Towards a genealogical classification of Taa dialects. In Tom Güldemann & Anne-Maria Fehn (eds.), *Beyond 'Khoisan': Historical relations in the Kalahari Basin*, 283–301. Amsterdam: John Benjamins.
- Naumann, Christfried. 2017. The phoneme inventory of Taa (West !Xoon dialect). In Rainer Voßen & Wilfrid H.G. Haacke (eds.), *Lone Tree scholarship in the service of the Koon: Essays in memory of Anthony T Traill*, 311–351. Köln: Rüdiger Köppe.
- Ono, Hitomi. 2010. |Gui kinship verbs? Verbs and nouns in |Gui and linguistic differences found among its kinship terms. In Matthias Brenzinger & Christa König (eds.), *Khoisan languages and linguistics: Proceedings of the 1st International Symposium*, January 4-8, 2003, Riezlern/Kleinwalsertal, 251–283. Köln: Rüdiger Köppe.
- Pratchett, Lee J. 2017. *Dialectal diversity in Southeastern Ju (Kx'a) and a documentation of Groot Laagte #Kx'aol'ae*. Berlin: Humboldt Universität Berlin dissertation.
- Rapold, Christian J. 2012. The encoding of placement and removal events in †Akhoe Hailom. In Anetta Kopecka & Bhuvana Narasimhan (eds.), *Events of putting and taking: A crosslinguistic perspective*, 79–96. Amsterdam: John Benjamins.
- Rapold, Christian J. & Thomas Widlok. 2008. Dimensions of variability in northern Khoekhoe language and culture. In Karim Sadr & François-Xavier Fauvelle-Aymar (eds.), *Khoekhoe and the earliest herders in southern Africa*. *Southern African Humanities* 20. 133-161.
- Schladt, Mathias (ed.). 1998. *Language, identity, and conceptualization among the Khoisan*. Köln: Rüdiger Köppe.
- Schladt, Mathias. 2000. A multi-purpose orthography for Kxoe. In Herman M. Batibo & Joe Tsonope (eds.), *The state of Khoesan languages in Botswana*, 125–139. Gaborone: Basarwa Languages Project.
- Schmidt, Sigrid (ed.). 2011. *Hailom and !Xú stories from North Namibia: Collected and translated by Terttu Heikkinen (1934-1988)*. Köln: Rüdiger Köppe.
- Shah, Sheena & Matthias Brenzinger (eds.). 2016. *Khoisan languages and linguistics: Proceedings of the 5th International Symposium*, July 13–17, 2014, Riezlern/Kleinwalsertal. Köln: Rüdiger Köppe.
- Snyman, Jan W. 1975. *Žul'hōasi fonologie and woordeboek*. Cape Town: A. A. Balkema.
- Snyman, Jan W. 1998. An official orthography for Žul'hōasi Kokx'oi. In Schladt, Mathias (ed.), *Language, identity, and conceptualization among the Khoisan*, 95–115. Köln: Rüdiger Köppe.
- Snyman, Jan W. 2000. Palatalisation in the Tsowaa and Gllana languages of central Botswana. In Herman M. Batibo & Joe Tsonope (eds.), *The state of Khoesan languages in Botswana*, 33–43. Gaborone: Basarwa Languages Project.
- Tanaka, Jiro & Sugawara, Kazuyoshi (eds.). 2010. *An encyclopedia of |Gui and !Gana culture and society*. Kyoto: Laboratory of Cultural Anthropology, Kyoto University.
- Traill, Anthony. 1985. *Phonetic and phonological studies of !Xóó Bushman*. Hamburg: Helmut Buske.
- Traill, Anthony (ed. Hirsosi Nakagawa & Andy Chebanne). 2018. A trilingual !Xóó dictionary. Köln: Rüdiger Köppe.
- Visser, Hessel. 2000. The Khoisan orthography revisited. In Herman M. Batibo & Joe Tsonope (eds.), *The state of Khoesan languages in Botswana*, 145–160. Gaborone: Basarwa Languages Project.

- Visser, Hessel. 2001. *Naro dictionary: Naro-English/ English-Naro*. Gantsi: Naro Language Project.
- Visser, Hessel. 2010. Verbal compunds in Naro. In Matthias Brenzinger & Christa König (eds.), *Khoisan languages and linguistics: Proceedings of the 1st International Symposium*, January 4-8, 2003, Riezlern/Kleinwalsertal, 176–200. Köln: Rüdiger Köppe.
- Vossen, Rainer (ed.). 2013. *The Khoesan languages*. London: Routledge.
- Widlok, Thomas. 2008. Landscape unbounded: space, place, and orientation in †Akhoe Haillom and beyond. *Language Sciences* 30. 362–380.
- Widlok, Thomas. 2016. Small words – big issues: The anthropological relevance of Khoesan interjections. *African Study Monographs*, suppl. 52. 135–145.
- Widlok, Thomas, Christian J. Rapold & Gertie Hoymann. 2008. Multimedia analysis in documentation projects: Kinship, interrogatives and reciprocals in †Akhoe Haillom. In K. David Harrison, David S. Rood & Arienne Dwyer (eds.), *A world of many voices: Lessons from documented endangered languages*, 355–370. Amsterdam: John Benjamins.
- Witzlack-Makarevich, Alena & Martina Ernszt (eds.). 2013. *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium*, July 6-10, 2008, Riezlern/Kleinwalsertal. Köln: Rüdiger Köppe.

Tom Güldemann
tom.gueldemann@staff.hu-berlin.de

From comparative descriptive linguistic fieldwork to documentary linguistic fieldwork in Ghana

Felix K. Ameka
Leiden University

This paper surveys linguistic fieldwork practices in Ghana from the earliest times to the documentary linguistic era. It demonstrates that the most profound effect of the documentary linguistic turn in the language sciences on fieldwork in Ghana is in the rise of “insider” and “insider-outsider” fieldworking linguists. This goes against the definition of prototypical fieldwork as something done by remote outsiders. The challenges and opportunities of this development are reflected upon. It is argued that relevant fieldwork methodologies should be further developed taking the emerging features of different “insider” practices into account. Moreover, it is hoped that characterizations of documentary linguistic fieldwork would move beyond the outsider and accommodate the different types of “insider” fieldworkers.

1. Introduction Perhaps the most profound impact of the documentary linguistics turn in the language sciences on the study of Ghanaian languages has been in fieldwork practices. Its effect has been two-fold. First, it has led to the shift in the practices from native speaker linguists investigating their own languages, based on a combination of introspective methods and interaction with other native speakers, to trained native speaker linguists engaging in documentary linguistic fieldwork: gathering data, archiving and analysis. Second, it has also led to an increase in a category of fieldworkers and language documenters that I call “insider-outsiders”. These are people who are “insiders” by virtue of the fact that they share and participate in the wider Ghanaian community. They identify as Ghanaians just like members of the community they research, and in most cases they have as part of their linguistic repertoires the languages of wider communication such as Ewe and Akan that form part of the mosaic of languages in the communities of speakers of the languages they document (cf. Anyidoho & Dakubu 2008). In their recent textbook, Felicity Meakins et al. (2018: 4) characterize the “insider-outsider” as someone from a marginalized group, or, if from the same country, from

another minority group, but not from the speech community being researched. As an example of the former, they mention Annelies Kusters, a deaf researcher from Belgium who works in the Adomorobe deaf community in Ghana as an “insider-outsider” as she belongs to the wider deaf culture, but lacks specifics of the Adomorobe (Ghanaian) context (Kusters 2012). In my understanding of the term, the “insider-outsider” researcher need not come from a minority group. What is critical is that they should have some knowledge about the wider cultural norms and practices of the community but they are not local members. These “insider-outsiders” are different from the “insiders” who are native speakers and members of the linguaculture being documented or researched. Note that in the prototypical conception of fieldwork (e.g., Bower 2008, Newman & Ratliff 2001, Sakel & Everett 2012), these “insiders” are not considered typical fieldworkers. Fieldwork is usually carried out by an “outsider” who comes to spend time in the researched community. This aspect of linguistic fieldwork may be a vestige of its roots in ethnography where there is no place for “insider” ethnographers for they are not foreign to the researched community. Yet “insiders” (of different types) have lots of insights that are not easily accessible to the “outsider” (see Owusu 1978 and references therein on the problems of evaluating outsider ethnographies by “insiders”).

In this paper I reflect on fieldwork practices especially as they have been shaped over the past twenty years. To appreciate the developments, I provide a historical overview of the different types of linguistic fieldwork that have been carried out in Ghana since the end of the 19th century. Then I survey the developments in the documentary linguistics era. I conclude with reflections on how to move documentary fieldwork in Ghana forward and expand on the available documentary corpora. In particular, I advocate that more native speakers of lesser studied languages in the country should be motivated to be trained as linguists (see also Ameka 2006a). Moreover, there should be more done to increase archived language materials from Ghanaian languages, from both the better studied languages and the under-studied ones.

2. The early years Prototypical linguistic fieldwork, with the goal of collecting primary data through interaction with native speakers in the natural habitat using multiple methods including observation and interview, has long been practiced with Ghanaian languages. Missionaries as well as colonial explorers sought to understand the ethnolinguistic groups among whom they were working through a study of the languages. Notable products from this era are authoritative and insightful grammars, dictionaries and text collections of three bigger languages of southern Ghana: Gã (Zimmermann 1858), Twi (Akan) (Christaller 1875, [1881] 1933) and Ewe (Westermann 1907, 1930, 1928, 1954). Around the same time there was fieldwork related to the collection of data for determining the genetic relations among the languages. Some of the people involved in addition to Dietrich Westermann (e.g., 1922) and Johan Gottlieb Christaller were Rudolf Plehn (1898), and Emil Funke (e.g. 1910, 1920).

Also of note in this period are two descriptions of the Fante variety of Akan. There was Balmer & Grant (1929) based on introspection who seemed to have been the first to use the term “Serial Verb Construction” in the description of a typical verb construction of West African languages. Another grammar of Fante was produced by Welmers (1946), who worked in the US with the first President of Ghana, Kwame Nkrumah, as the language expert. Kwame Nkrumah was bilingual in Nzema and Fante. Their interaction represents an example of descriptive fieldwork outside the linguistic milieu and in a university office environment (see Hyman 2002). It also exposes the extent to which linguists in those

days paid little attention to the linguistic life histories and repertoires of their consultants. Kwame Nkrumah, from all accounts, was Nzema dominant. Hence the grammar could be described as a grammar of a second language speaker of Fante.

3. The last half of the twentieth century

3.1 SIL International and GILLBT A landmark in the fieldwork-based investigations of lesser studied Ghanaian languages was the beginning of work in the country by SIL International in 1962. SIL entered a partnership with the newly established Institute of African Studies (IAS) at the University of Ghana, Legon. The first SIL-Ghana Director, John Bendor-Samuel, was a Scientific Staff member of IAS. SIL sent teams to different communities. They published a series of Collected Language Notes at IAS. These were field reports containing wordlists and grammatical notes gathered *in situ*. Some of these have remained the only sources on several Ghanaian languages to date (e.g., Crouch & Smiles 1966 on Vagala). The goal of these activities was literacy development leading to Bible Translation. Some teams have conducted sociolinguistic surveys as well (e.g. Ring 1981). When Ghana SIL became autonomous, its name changed to the Ghana Institute of Linguistics, Literacy and Bible Translation (GILLBT) with headquarters in Tamale in the northern part of the country. Descriptive linguistic work by GILLBT teams is primarily carried out by “outsiders”. The documentation of Nkonya by Wesley Peacock with its online dictionary is exemplary.¹ GILLBT members have continued the tradition of the Collected Language Notes and published more substantial descriptions of some of the languages in the Language Monograph Series also published by IAS (see e.g., Casali 1995, Smye 2004).

Among the output of the GILLBT teams are also several dissertations which provide a comprehensive description of hitherto undocumented languages based on immersion fieldwork. These include works on Kɔnni (Michael Cahill 2007), Chumburung (Keir Hansford 1990), Tuwuli (Matthew Harley 2005) and Safaliba (Paul Schaefer 2009). Works on semantics and pragmatics, as well as anthropological topics, have also been carried out by some GILLBT members, e.g., Tony Naden (1986, 1993) and the late Gillian Hansford (2005, 2012). Although there are more Ghanaians working with GILLBT these days, I am aware of only one “insider-outsider” linguistic work on the Ahanta language (Ntummy 2002).

In addition, GILLBT teams are involved in dictionary work on some of the languages. Apart from earlier published dictionaries such as Blass (1975), there is a major dictionary project on Adele, a Ghana-Togo Mountain (GTM) language. There is also an ongoing Buli (Gur, Mabia) dictionary for which the Rapid Word Collection Method has been used.²

While many of the publications of the GILLBT teams are available, the extensive recorded data collections on which they are based are less available and accessible.

3.2 The era of the West African Languages Survey and beyond In the early 1960s, another major project—the West African Languages Survey—initiated and directed by Joseph H. Greenberg, was launched in Ghana. The goal was to gather data through interactions with speakers of the languages for comparative-historical studies, as well as for uncovering typological features of the languages. Ladefoged’s (1964) *Phonetic*

¹<https://nkonya.webonary.org/>

²<https://vimeo.com/44131617>

study of West African languages was one of the outcomes of the project with significant contributions on Ghanaian languages. This is a pioneering work in phonetic fieldwork. The recordings are archived and available.³ Indeed, Ghanaian languages have contributed to the development of fieldwork techniques for investigating the phonetics of tone languages (see e.g., Gleason 1961 on Ewe). This tradition was later followed in the 1990s by Russell Schuh (e.g., 1995) and Ian Maddieson (e.g., 1998), who investigated some GTM languages. Another outcome of the West African Languages Survey with relevance for Ghanaian languages is the West African Languages Data Sheets, which has grammatical notes and vocabularies from several languages (Dakubu 1977, 1980)

The late 1960s and early 1970s saw intense fieldwork activity mainly by “outsider” linguists. One of the foci at this time was the collection of wordlists for comparative historical linguistics studies. Thus, at the IAS several wordlists for various languages were collected and published (e.g., Stewart 1966, Kropp 1967, and later Snider 1989). The source items for these lists have found their way into the SIL African Comparative Wordlist (Snider & Roberts 2004). At this time also, fieldwork for the seminal comparative historical study of the GTM languages (Heine 1968) was carried out. The other strand of work was related to descriptions of the languages informed by the linguistic models of the time, especially the generative paradigm as outlined in Chomsky & Halle’s (1968) *Sound Patterns of English* (for phonology) and Chomsky’s (1965) *Aspects of the theory of syntax* (for syntax). At this time, the first PhD degree to be awarded in linguistics by the University of Ghana was bestowed on Kevin Ford for his description of *Aspects of Avatime syntax* (Ford 1971), a GTM language. Another dissertation based on extensive fieldwork was produced by Alan (1971) on Buem, another GTM language. The late Nick Clements also conducted fieldwork on the Anlo (Anyako) dialect of Ewe, leading to a dissertation on the verbal syntax of Ewe (Clements 1972). This work is important as it is a precursor to current micro-variation studies of syntax. The Humboldt University of Berlin carried out field investigations on the languages of the Central Togoland, generating grammatical descriptions of Nkonya, a Guang language (Reinecke 1972), and Lelemi (Buem, GTM; Höftmann 1971) and a collection of oral traditions Höftmann & Ayitevi (1968). At this time Colin Painter also carried out a survey of Northern Guang languages and produced a grammatical description of Gonja (Painter 1967, 1970).

The contribution of the late Mary Esther Kropp Dakubu to field investigations of Ghanaian languages is phenomenal. She carried out comparative historical work to reconstruct the Gã-Dangme subgroup of Kwa. She did ethnographic fieldwork on a Gã clan (Dakubu 1981). In addition, she compiled dictionaries of Gã and Dangme (Dakubu 1999). She also published a grammar of Dangme (Dakubu 1987) and a phonological description of Gã (Dakubu 2002). She has collected grammatical notes on other languages such as Dagaare, and has collaborated with native speakers of Gurene to produce its first dictionary (Dakubu et al. 2007). She also published a description of the language and culture of the Gurene people (Dakubu 2009) based on ethnographic fieldwork.

Sociolinguistic investigations based on structured interviews and questionnaires were also carried out at this time. There was the IAS Madina Project, the aim of which was to study the linguistic situation and profiles of members of a new suburb of Accra. Surveys of language use in markets across Ghana were conducted for language planning purposes.

Dialect surveys were also conducted at this time. A significant example is the study of the cluster of languages Ewe-Gen-Aja-Fon (now called Gbe) proposed by Capo in his

³<http://archive.phonetics.ucla.edu/>

University of Ghana 1981 dissertation that was published in 1991. Similar comparisons of the dialects of Akan were undertaken by Lawrence Boadi (e.g., 2009) and Florence Dolphyne (e.g., 1976). Even though various works on varieties of English had been carried out since the 1960s, it was only in the mid 1990s that some field investigations were carried out especially on Ghanaian Pidgin English. The description by Magnus Huber (1999) contains a CD of the field recordings upon which the book is based. It was the first work to make available, publicly, the field recordings in a publication on Ghanaian languages.

3.3 From outsider descriptions to native speaker descriptions Native speaker linguists took the stage during the era of theoretical linguistics. This era coincides also with the West African Languages Survey outlined in the previous section. With the rise of theoretical linguistics boosted by Chomsky's (1957) *Syntactic Structures*, there was a particular interest in a trained native speaker as the linguist *par excellence*. Such persons were considered the "ideal native speakers". The early 1960s saw the emergence of investigations of major Ghanaian languages by such linguists using various theoretical models of the day: generative linguistics, Firthian linguistics and Halliday's Systemic Functional Grammar or Scale and Category Grammar. The main method at this time was introspection based on the idiolect and dialect of the native speaker researcher.

The first description of a Ghanaian language based on modern linguistic principles authored by a native speaker is the seminal study of Ewe tonal structure by Gilbert Ansre (1961). This was followed by his doctoral dissertation on the grammatical units of Ewe using Halliday's Scale and Category framework (Ansre 1966). Similar foundational descriptions of Akan authored by trained native speaker linguists were by Florence Dolphyne (1965) on phonetics and Lawrence Boadi (1966) on the generative syntax of Twi. Other native speaker works using themselves or their family as consultants include Isaac Chinebuah's (1963) Master's thesis on the phonology of Nzema, based on his own pronunciation. Similarly, the very first studies of child language in a Ghanaian language is based on Eric Apronti's observations of his 2 year old child (Apronti 1969). Issues of contact between English and the indigenous languages gave rise to studies of code-switching by native speakers (e.g., Forson 1979, Amuzu 2005).

The next generation of native speaker linguists working on their own languages augmented introspection with interviews of other native speakers for acceptability judgments and interpretations. They also relied less on their own production and, for the languages with some literature, illustrative examples were drawn from novels, plays and newspapers. Others also used media broadcasts such as radio and television drama. Thus, data sources expanded, and constructed examples based on introspection minimized. This is the mark of doctoral dissertations such as Osam (1994) and Saah (1994) on Akan; Bodomu (1997) on Dagaare; and Ameka (1991) and Essegbey (1999) on Ewe. In fact, Essegbey (1999) is probably a pioneer in using stimulus-based elicitation methods (Majid 2012) in gathering data as a native speaker (see also Kita & Essegbey 2001).

4. Language documentation and description in the 21st Century

4.1 Insider fieldwork In the 2000s, research on Ghanaian languages has continued the earlier traditions of descriptions by native speakers based on introspection and interaction with other speakers, plus the use of existing written literature and language use in mass

media as sources of examples. They can be considered as studies based on one or the other form of fieldwork by “insiders”. Many of these descriptions are dissertations. Recent descriptions of Gur/Mabia languages following this tradition include Musah (2018) and Abubakar (2018) on Kusaal, Mwinlaaru (2017) on Dagaare and Hudu (2010) on Dagbani. For Kwa languages, Akrofi-Ansah (2009) on *Letɛ*, Agyepong (2017) investigated ‘cutting’ and ‘breaking’ events in Asante Twi (Akan), Campbell (2017) is a comprehensive grammar of *Gã*, Abunya (2018) is an examination of aspects of *Kaakye* syntax and there is on-going dissertation project on *Gwa* by Michael Obiri-Yeboah.

Other native speakers have embarked on documentary fieldwork in their own communities; recording and analyzing different genres of language use. The products are mainly dissertations. Some of these have archived their data. Examples are Cephas Delalorm (2016) on *Likpe* and Samuel Atintono (2014) on *Gurene*. These primary data are accessible at ELAR. Similarly Dodzi Kpoglu (2019) collected *Tɔɔɔgbe* primary data to investigate possession and the material available through DANS.⁴

A challenge with “insider” fieldwork is the suspicion with which consultants look on the researcher. As a community member the “insider” researcher is expected to share the same knowledge and practices being investigated. Why then do the researchers interview them about this same knowledge?

4.2 Fieldwork by insider-outsiders Historically, “insider-outsider” descriptions came later than works by “insiders”. Samuel Obeng for example, worked on languages other than his first language Akan. He conducted descriptive linguistic fieldwork on the *Guang* language *Efutu* produced a grammar (Obeng 2008).

The era of “insider-outsider” documentary fieldwork came to be established starting from my work in the mid 1990s on *Likpe* (*Sɛkpelé*). It gained ascendancy with the launch of the Southern Ghana-Togo Mountain languages (SGTM) project in 2003 funded by the Netherlands Science Foundation (NWO) under its Endangered Languages Programme (ELP). The researchers in this project are James Essegbey (*Nyagbo/Tutrugbu*, 2019), Kofi Dorvlo (*Logba/Ikpana*, 2008) and Mercy Bobuafor (*Tafi*, 2013). They are all non-native speakers of the individual languages they work on. They all have *Ewe*, the major lingua franca in the communities, as their primary language. Essegbey also works on *Dwang* and *Animere*, communities that use *Twi* as lingua franca, another of his primary languages. Since then, many more researchers have embarked on documentary fieldwork in communities where they are “insider-outsiders”. They include Rogers Asante (2016) on *Nkami* and Nana Ama Agyeman (2014) on *Efutu* who have archived their materials at ELAR, as well as Yvonne Agbetsoamedo (2014) on *Seleɛ* and other on-going dissertation work by Esther Dogbe on *Dompo*, a highly endangered, presumably isolate, language of the middle belt. Obed Nii Broohm and Victoria Owusu are also working on *Esahie*, a Central Tano Kwa language.

While “insiders” are viewed with suspicion for not knowing their own lingua-cultures, the “insider-outsiders” are suspected of different things. In my own case, some expressed doubt about why I would want to learn their language. Some thought it must be the case that I was dating one of their women and I wanted to impress her by learning the language. In the case of others, they are viewed with suspicion and skepticism. One researcher was viewed as probably being a government agent or a private detective, especially because he had recording gear. Some others get caught up in internal conflicts and are viewed as siding

⁴<http://doi.org/10.17026/dans-xxr-4sug>

with one or the other party (cf. Essegbey in press). Yet others are warmly welcomed and seen as facilitators of empowerment. Rogers Asante reports how the Nkami people saw his work as a way of getting their language written, thereby removing a stigma and source of insults that the neighbouring groups hurl at them. “Insider-outsider” research is on the increase and the future lies in more collaboration between these and trained “insiders” of various communities for providing long-lasting records of the many under-documented languages of Ghana.

4.3 Fieldwork by outsiders Alongside ‘insiders’ as well as “insider-outsiders”, the classical tradition of fieldwork by complete “outsiders” has continued in the documentary era. The main difference between earlier “outsider” research and current practice is the increased use of audio-visual recordings, accountability of data, and above all the accessibility of data collections on the various languages. There is also an increased sense of ‘giving back to the community’ among the modern day “outsider” researchers. One of the first documentation projects awarded by the Endangered Languages Documentation Programme (ELDP) was for the documentation of Cala (Kleinwillinghoffer 2007). Fieldwork on Avatime (GTM) by Saskia van Putten (2014) and Rebecca Defina (2016) was initially funded by ELDP and later the Max Planck Society. The same organization funded work on Siwu by Mark Dingemans (2011). Other “outsiders” working on Ghanaian languages include Victoria Nyst (2007) on Adamorobe Sign Language, Jonathan Brindle (2017) on Chakali and Cedar (formerly Lydia) Green (2009) on Logba ethnobotany. Purvis (2008) on Dagbani, Jason Kandybowicz and Harold Torrence (e.g., 2017) have also been conducting syntactic fieldwork on Kaakye (Guang, Kwa). These “outsiders” have striven to work with the communities managing speaker ideologies and attitudes in obtaining optimal records of the languages. They have been well accepted into the communities with one of them being formally initiated into a sacred society.

5. Going forward The conceptions of fieldwork in the standard manuals (e.g., Chelliah & de Reuse 2011, Bovern 2008, Sakel and Everett 2012, Dixon 2010) assume that fieldwork involves an “outsider”. Aikhenvald (2007: 5) indicates that it involves an outsider “becoming a member of the community and often becoming adopted into the kinship system”. It is not surprising that linguistic fieldwork with its roots in ethnographic and anthropological fieldwork should consider fieldwork to be prototypically and usually carried out by “outsiders”. Nevertheless, it is a paradox in the documentary linguistics era for “outsider” fieldwork to be considered the norm. Chelliah & de Reuse (2011: 10–11) suggest that fieldwork in the documentary era involves “collection or gathering of linguistic data through a variety of methods and techniques with a focus on reliability, representativity and archivability” (cf. Thieberger 2012). Given such a view it would appear that fieldwork does not have to be an enterprise of an outsider to the community whose “doculect” is being investigated. Moreover, if one of the canons of documentation is that of collaboration with the community, what would be better than having professionally trained community members being the researchers of their own communities (cf. Newman 2003)? The only hope of having many more languages of Ghana documented is when more trained linguists engage in such work rather than focusing on their languages and modeling them in the latest fashionable yet ephemeral formal architecture.

There is a second aspect to the received practice of fieldwork that is not conducive to native speakers of African languages being the researchers. Newman & Ratliff (2001)

in their reflections on linguistic fieldwork note that their book and its discussions of fieldwork methodologies is constrained and focused on practices developed primarily by North American, Australian and Western European linguists. It is my hope that the practice of “insider” and “insider-outsider” documentation and fieldwork whose beginnings we see in Ghana and elsewhere in Africa will impact fieldwork methodologies. With a third of the world’s languages spoken in Africa, methodologies and techniques developed and lessons learned in Africa by trained professional community members should also feed into the global discourse about documentary linguistic fieldwork.

Table 1: Ghanaian languages with documentary materials

Language	ISO	Type of researcher	Output	Institution
Adamorobe Sign Language (AdaSL)	ads	Outsider, Victoria Nyst (2007)	PhD thesis, grammar	University of Amsterdam/Leiden University repository
Animere	anf	Insider-outsider, outsiders (in progress), James Essegbey, Bryan Gelles	Multi-media record of linguistic practices	ELDP/ELAR, DEL-NSF University of Florida, Gainesville
Avatime (Siya)	avm	Outsiders, Defina (2016), van Putten (2014)	Archived documentary corpora MA/PhD theses on information structure; event description	ELDP/ELAR, The Language Archive (MPI, Nijmegen)
Chala (Bogon)	cll	Outsider, Kleinewillinghöfer (2007)	Archived documentary corpus	ELDP/ELAR
Chakali	cli	Outsider, Brindle (2012, 2017)	PhD thesis, grammar; published dictionary	Norwegian University of Technology, Trondheim
Dompo	doy	Insider-outsider, Esther Dogbe, Outsider, Roger Blench	PhD thesis in progress; sociolinguistics, grammar and texts wordlists	La Trobe University, Melbourne Australia
Dwang	nnu	Insider-outsider, James Essegbey	Archived multi-media record; thematic documentation of fishing practices	ELDP/ELAR

Continued on next page

Table 1 – *Continued from previous page*

Language	ISO	Type of researcher	Output	Institution
Effutu	afu	Insider-outsiders, Nana Ama Agyemang, Samuel Obeng	Archived multi-media record; thematic grammatical description descriptive grammar	ELDP/ELAR
Esahie (Sehwi)	sfw	Insider-outsiders, Obed Nii Broohm, Victoria Owusu	PhD theses on different aspects of grammar (in progress)	University of Verona; University of Ghana
Farefare/ Gurene	gur	Outsider collaborating with insiders, Dakubu et al. (2009), Samuel Atintono	Dictionary, a documentation of Gurene folk tales, riddles, songs, palace genres and other oral genres in Bolga	ELDP/ELAR University of Manchester
Gwa	gwx	Native speaker Michael Obiri-Yeboah	PhD research (on going)	University of California, San Diego
Kaache/ Krachi/ Kaakyi	kye	Outsiders, Jason Kandybowicz & Harold Torrence (2017), Levina Abunya (2018)	Description, topics, e.g., Questions PhD thesis, grammatical description	NSF, University of Ghana
Logba (Ikpana)	lgq	Insider-outsider, Dorvlo (2008, 2011), outsider Cedar (formerly Lydia) Green (2009)	PhD thesis; grammar, texts, dictionary language use in Logba schools ethno-botanical documentation of plant names	Leiden University University of Ghana
Nyagbo/ Tutrugbu	nyb	Insider-outsider, Essegbey (2019)	Archived multi-media recordings reference grammar, dictionary, texts	The Language Archive (MPI, Nijmegen)

Continued on next page

Table 1 – Continued from previous page

Language	ISO	Type of researcher	Output	Institution
Sekpele	lip	Insider (native speaker), Delalorm (2016), Insider-outsider, Ameka (2006b, 2007, 2009)	Archived multi-media recordings PhD thesis, grammar, texts thematic documentation of semantic categories (e.g., space)	ELDP/ELAR School of Oriental and African Studies, University of London
Selee	snw	Insider-outsider, Agbetsoamedo (2014)	PhD thesis, aspects of grammar and lexicon	University of Stockholm
Siwu	akp	Outsider, Dingemanse (2011) Outsider and insider linguists: Ford & Idah (2017)	Multi-media archived recordings PhD thesis on ideophones (Siwu Akpafu); reference grammar on Siwu-Lolobi	The Language Archive (MPI, Nijmegen)
Tafi	tcd	Insider-outsider, Bobuafor (2013)	PhD thesis, grammar and texts	Leiden University
Tuwuli	bov	Outsider, Harley (2005)	PhD thesis, grammar	School of Oriental and African Studies, University of London; GILBLLT
Women's language Kiliji	—	Outsiders with collaboration of outsider-insiders	Recordings and annotated texts including songs	ELDP/ELAR Institut of African Studies, University of Ghana, Legon Brindle et al. (2015)

References

- Abubakari, Hasiyatu. 2018. Aspects of Kusaal grammar: The syntax-information structure interface. Vienna: University of Vienna dissertation.
- Abunya, Levina. 2018. Aspects of Kaakye syntax. Accra: University of Ghana, Legon dissertation.
- Agbetsoamedo, Yvonne. 2014. Aspects of the grammar and lexicon of Sɛlɛɛ. Stockholm: University of Stockholm dissertation.
- Agyeman, Nana Ama. 2014. Verb serialization in Efutu. London: School of Oriental and African Studies, University of London dissertation.
- Aikhenvald, Alexandra Y. 2007. Linguistic fieldwork: Setting the scene. *STUF—Sprachtypologie und Universalienforschung* 60(1). 3–11.
- Akrofi Ansah, Mercy. 2009. *Aspects of Lɛtɛ grammar*. Manchester: University of Manchester dissertation.
- Allan, Edward Jay. 1973. *A grammar of Buem, the Lelemi language*. London: University of London dissertation.
- Ameka, Felix K. 1991. *Ewe: Its grammatical constructions and illocutionary devices*. Canberra: Australia National University dissertation.
- Ameka, Felix K. 2006a. Real descriptions: Reflections on native speaker and non-native speaker descriptions of a language. In Felix K. Ameka, Alan Dench & Nicholas Evans (eds.) *Catching language: the standing challenge of grammar writing*, 70–112. Berlin: Mouton de Gruyter.
- Ameka, Felix K. 2006b. Grammars in contact in the Volta Basin (West Africa): On contact induced grammatical change in Likpe. In Alexandra Y. Aikhenvald & R.M.W. Dixon (eds.) *Grammars in contact: A cross-linguistic typology*, 114–142. Oxford: Oxford University Press.
- Ameka, Felix K. 2007. The coding of topological relations in verbs: the case of Likpe (Sɛkpeɛlɛ). *Linguistics* 45(5/6). 1065–1103.
- Ameka, Felix K. 2009. Likpe. In Gerrit J. Dimmendaal (ed.) *Coding participant marking construction types in twelve African languages*, 239–280. Amsterdam: John Benjamins.
- Amuzu, Evershed Kwasi. 2005. Ewe-English codeswitching: A case of composite rather than classic codeswitching. Canberra: Australia National University dissertation.
- Anyidoho, Akosua & Mary Esther Kropp Dakubu. 2008. Ghana: Indigenous languages, English, and an emerging national identity. In Andrew Simpson (ed.) *Language and National Identity in Africa*, 141–157. Oxford: Oxford University Press.
- Apronti, Eric O. 1969. The language of a two-year old Dangme. *Proceedings of the 8th West African Languages Conference*, 19–29. Abidjan: Institut de Linguistique Applique.
- Asante, Rogers Krobea 2016. Nkami language: Description and analysis. Shanghai: Tongji University dissertation.
- Atintono, Awinkene Samuel. 2014. The semantics and grammar of Gureɛ positional verbs: A typological perspective. Manchester: University of Manchester dissertation.
- Ansre, Gilbert. 1961. *The tonal structure of Ewe*. Connecticut: Hartford Seminary Foundation.
- Ansre, Gilbert. 1966. The grammatical units of Ewe. London: School of Oriental and African Studies, University of London dissertation.
- Balmer, William Turnbull & F. C. F. Grant. 1929. *A grammar of the Akan-Fante language*. London: Atlantis Press.

- Blass, Regina. (ed). 1975. *Sisaala-English, English-Sisaala dictionary*. Tamale: Ghana Institute of Linguistics, Literacy and Bible Translation.
- Boadi, Lawrence. 1966. The syntax of the Twi verb. London: University of London dissertation.
- Boadi, Lawrence. 2009. *A comparative phonological study of some verbal affixes in seven Volta Comoe languages of Ghana*. Accra: Black Mask.
- Bobuafor, Mercy. 2013. A grammar of Tafi. Leiden: Leiden University dissertation.
- Bodomo, Adams. 1997. Paths and pathfinders: Exploring the syntax and semantics of complex verbal predicates in Dagaare and other languages. Trondheim: Norwegian University of Science and Technology.
- Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. New York: Palgrave Macmillan.
- Brindle, Jonathan A. 2012. Aspects of the Chakali language. Trondheim: Norwegian University of Science and Technology dissertation.
- Brindle, Jonathan. 2017. *A dictionary and grammatical outline of Chakali*. Berlin: Language Science Press.
- Brindle, Jonathan, Mary Esther Kropp Dakubu & Ọbádélé Kambon. 2015. Kiliji, an unrecorded spiritual language of Eastern Ghana. *Journal of West African Languages XLII* 1. 65–88
- Cahill, Michael, 2007 *Aspects of the morphology and phonology of Konni*. Dallas: SIL International and the University of Texas at Arlington Publications in Linguistics.
- Campbell, Akua. 2017. *A grammar of Gã*. Houston: Rice University dissertation.
- Capo, Hounkpati B. C. 1991. *A comparative phonology of Gbe*. Berlin: Mouton.
- Casali, Roderic. 1995. *Nawuri phonology*. Legon: Institute of African Studies, University of Ghana.
- Chelliah, Shobhana L., & J. Willem de Reuse. 2010. *Handbook of descriptive linguistic fieldwork*. Dordrecht: Springer Science & Business Media.
- Chinebuah, Isaac Kodwo. 1963. A phonetic and phonological study of the nominal piece in Nzema, based on the candidate's own pronunciation. London: School of Oriental and African Studies, University of London dissertation.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.
- Chomsky, Noam & Morris Halle. 1968. *The sound patterns of English*. New York: Harper & Row.
- Christaller, Johann Gottlieb. 1875. *A grammar of the Asante and Fante language called Tshi Chwee, Twi based on the Akuapem dialect with reference to the other (Akan and Fante) dialects*. Basel: Evangelical Missionary Society.
- Christaller, Johann Gottlieb. 1933[1888]. *A dictionary of the Asante and Fante language called Tshi (Chwee, Twi): With a grammatical introduction and appendices on the geography of the Gold Coast and other subjects*. Basel: Evangelical Missionary Society.
- Clements, G. Nick. 1972. The verbal syntax of Ewe. London: University of London dissertation.
- Crouch, Marjorie & Nancy Smiles. 1966. *Collected field reports on the phonology of Vagala*. Accra: Institute of African Studies, University of Ghana.
- Dakubu, Mary Esther Kropp (eds.). 1977. *West African language data sheets, Volume 1*. Accra: West African Linguistic Society.

- Dakubu, Mary Esther Kropp (eds.) 1980. *West African language data sheets, Volume 2*. Leiden: African Studies Centre.
- Dakubu, Mary Esther Kropp. 1981. *One voice: The linguistic culture of an Accra lineage*. Leiden: African Studies Centre.
- Dakubu, Mary Esther Kropp 1987. *The Dangme language: An introductory Survey*. London: Macmillan.
- Dakubu, Mary Esther Kropp, (ed.) 1999. *Ga-English dictionary with English-Ga index*. Accra: Black Mask.
- Dakubu, Mary Esther Kropp. 2002. *Ga phonology*. Legon: Institute of African Studies, University of Ghana.
- Dakubu, Mary Esther Kropp. 2009 *Parlons farefari (gurenè): Langue et culture de Bolgatanga (Ghana) et ses environs*. Paris: Editions L'Harmattan.
- Dakubu, Mary Esther Kropp, Samuel Awinkene Atintono & Ephraim Avea Nsoh. 2007. *Gurene-English dictionary: Gurene-English dictionary, Volume 1*. Legon: University of Ghana.
- Defina, Rebecca. 2016. *Events in language and thought: The case of serial verb constructions in Avatime*. Nijmegen: Raboud University dissertation.
- Delalorm, Cephas. 2016. *A grammar of Sekpele*. London: School of Oriental and African Studies, University of London dissertation.
- Dingemans, Mark. 2011. *The meaning and use of ideophones in Siwu*. Nijmegen: Raboud University dissertation.
- Dixon, R.M.W. 2010. *Basic linguistic theory*. Vol. 1. Oxford: Oxford University Press.
- Dolphyne, Florence Abena. 1976. Dialect differences and historical processes in Akan. *Legon Journal of the Humanities* 2. 15–27.
- Dolphyne, Florence Abena. 1965. The phonetics and phonology of the verbal piece in the Asante dialect of Twi. London: School of Oriental and African Studies, University of London dissertation.
- Dorvlo, Kofi. 2008. A grammar of Logba (Ikpana). Leiden: Leiden University dissertation.
- Dorvlo, Kofi. 2011. *Logba, English, Ewe Dictionary with English Logba index* Accra: School of Communications Press.
- Essegbey, James. 1999. Inherent complement verbs revisited: Argument structure constructions in Ewe. Leiden: Leiden University dissertation.
- Essegbey, James. 2019. *Tutrugbu (Nyagbo) language and culture*. Leiden: Brill.
- Essegbey, James. In press. Oral traditions and linguistic reconstructions. In Muaka, Leonard & Dainess Maganda (eds.), *Language, Literature, Education, and Liberation in Africa and the African Diaspora, Proceedings of the 7th SEALLF Conference*.
- Essegbey, James & Sotaro Kita. 2001. Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. *Gesture* 1. 73–95.
- Ford, Kevin C. 1971. Aspects of Avatime syntax. Accra: University of Ghana dissertation.
- Forson, Barnabas. 1979. Code-switching in Akan-English bilingualism. Los Angeles: University of California, Los Angeles dissertation.
- Funke, Emil. 1910. Die Nyangbo-Tafi Sprache: Ein Beitrag zur Kenntnis der Sprachen Togos. *Mitteilungen des Seminars für Orientalische Sprachen* 13(3). 166–201.
- Funke, Emil. 1920. Originaltexte aus den Klassensprachen in Mittel-Togo. *Zeitschrift für Eingeborenen Sprachen* 10. 261–313.
- Gleason, Henry Allan. 1961. *An Introduction to Descriptive Linguistics, Revised edition*. New York: Holt, Rinehart and Winston.


- Green, Lydia Jewl. 2009. A preliminary linguistic analysis of plant names in Ikpána (Logba), an endangered Ghana-Togo Mountain language. Independent Study Project (ISP) Collection. Paper 751. (http://digitalcollections.sit.edu/isp_collection/751).
- Hansford, Gillian F. 2005. My eyes are red: Body metaphor in Chumburung. *Journal of West African Languages* XXXII 1/2. 135–180.
- Hansford, Gillian F. 2012. Numbers that Chumburung people count on. In Anna Idström & Elisabeth Piirainen, in cooperation with Tiber F.M. Falzett (eds.) *Endangered metaphors* 221–252. Amsterdam: John Benjamins.
- Hansford, Keir Lewis. 1990. A grammar of Chumburung: A structure-function hierarchical description of the syntax of a Ghanaian language. London: School of Oriental and African Studies, University of London dissertation.
- Harley, Matthew. 2005. *A descriptive grammar of Tuwuli, a Kwa language of Ghana*. London: School of Oriental and African Studies, University of London dissertation.
- Heine, Bernd. 1968. *Die Verbreitung und Gliederung der Togo- und Togorestsprachen*. Berlin: Dietrich Reimer.
- Höftmann, Hildegard. 1971. *The structure of Lelemi language*. Leipzig: VEB Verlag Enzyklopädie.
- Höftmann, Hildegard & John K. Ayitevi. 1968. Oral traditions of some of the so-called remnant people in the Central Volta Region of Ghana. *Mitteilungen des Institut für Orientalforschung* 12. 199–219.
- Hyman, Larry M. 2001. Fieldwork as a state of mind. In Paul Newman & Martha Ratliff (eds.), *Linguistic Fieldwork*, 15–33. Cambridge University Press.
- Huber, Magnus. 1999. *Ghanaian Pidgin English in its West African context: A sociohistorical and structural analysis*. Amsterdam: John Benjamins.
- Hudu, Fusheini Angulu. 2010. *Dagbani tongue-root harmony: a formal account with ultrasound investigation*. Ph.D. Dissertation, University of British Columbia.
- Kandybowicz, Jason & Harold Torrence. 2017. The role of Theory in Documentation: Intervention effects and missing gaps in the Krachi documentary record. In Jason Kandybowicz & Harold Torrence (eds.), *Africa's Endangered Languages: Documentary and Theoretical Approaches*, 187–205. New York: Oxford University Press.
- Kleinewillinghöfer, Ulrich. 2007. *Bogoŋ aduuna na bɪndɛ atawɪsa: A collection of proverbs and wise sayings of the Chala people*. Legon: Institute of African Studies, University of Ghana.
- Kpoglu, Dodzi P. 2019. Possessive constructions in Tongugbe, a dialect of Ewe. Utrecht: Leiden University dissertation.
- Kropp, Mary Esther. 1967. *Lefana, Akpafu and Avatime with English gloss*. (Comparative African Wordlists no. 3.) Legon: Institute of African Studies, University of Ghana.
- Kusters, Annelies. 2012. Being a deaf white anthropologist in Adamorobe: Some ethical and methodological issues. In Ulrike Zeshke & Connie De Vos (eds.), *Sign languages in village communities: Anthropological and linguistic insights*, Vol. 4, 27–52. Berlin: Walter de Gruyter.
- Ladefoged, Peter. 1964. *A phonetic study of West African languages: an auditory-instrumental survey*. (West African Language Monograph Series 1.) Cambridge: University Press in association with the West African Languages Survey.

- Maddieson, Ian. 1998. Collapsing vowel harmony and doubly-articulated fricatives: two myths about the phonology of Avatime. In Ian Maddieson & Thomas J. Hinnebusch (eds.), *Language History and Linguistic Description in Africa*, 155–166. (Trends in African Linguistics 2.) Trenton: Africa World Press.
- Majid, Asifa. 2012. A guide to stimulus-based elicitation for semantic categories. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 54–71. Oxford University Press.
- Meakins, Felicity, Jennifer Green & Myfany Turpin. 2018. *Understanding linguistic fieldwork*. New York: Routledge.
- Musah, Agoswin. 2018. *A grammar of Kusaal, a Mabia (Gur) language, northern Ghana*. Berlin: Peter Lang.
- Mwinlaaru, Isaac Nuokyaa-Ire. 2017. *A systemic functional description of the grammar of Dagaare*. Ph.D. Dissertation, The Hong Kong Polytechnic University.
- Naden, Tony. 1986. Social context and Mampruli greetings. In George Huttaar (ed.), *Pragmatics in non-Western perspective*, 61–199. Arlington: University of Texas Press.
- Naden, Tony. 1993. From wordlist to comparative lexicography: the lexinotes. *Lexikos* 3. 167–190.
- Newman, Paul. 2003. The endangered languages issue as a hopeless cause. In Jansen, Mark & Sijmen Tol (eds). *Language Death and Language Maintenance: Theoretical, practical and descriptive approaches*, 1–14. Amsterdam: John Benjamins.
- Newman, Paul & Martha Ratliff (eds.). 2001. *Linguistic Fieldwork*. Cambridge University Press.
- Ntumy, Samuel K. 2002. *Collected field reports on the phonology of Ahanta*. No. Legon: Institute of African Studies, University of Ghana.
- Nyst, Victoria. 2007. *A descriptive analysis of Adamorobe Sign Language (Ghana)*. Utrecht: LOT
- Obeng, Samuel Gyasi. 2008. *Efutu grammar*. Lincom Europa.
- Olawsky, Knut J. 1999. *Aspects of Dagbani grammar: with special emphasis on phonology and morphology*. Munich: Lincom Europa.
- Osam, E. Kweku. 1994. *Aspects of Akan grammar*. Eugene: University of Oregon dissertation.
- Owusu, Maxwell. 1978. Ethnography of Africa: the usefulness of the useless. *American Anthropologist* 80(2). 310–334.
- Painter, Colin. 1967. The distribution of Guang in Ghana and a statistical pre-testing on twenty-five idiolects. *Journal of West African Languages* 4(1). 25–78.
- Painter, Colin. 1970. *Gonja: a phonological and grammatical study*. Bloomington: Indiana University Press.
- Peacock, Wesley. 2007. *The phonology of Nkonya*. (Collected Language Notes No. 27.) Legon: Institute of African Studies, University of Ghana.
- Plehn, Rudolf. 1898. Beiträge zur Völkerkunde des Togo-Gebietes. *Mitteilungen des Seminars für Orientalische Sprachen* 2, Part III, 87–124.
- Purvis, Tristan Michael. 2008. *A linguistic and discursive analysis of register variation in Dagbani*. Bloomington: Indiana University.
- Reineke, Brigitte. 1972. *The structure of the Nkonya language: with texts and glossary*. Leipzig: VEB Verlag Enzyklopädie,
- Ring, J. Andrew. 1981. Ewe as a second language: a sociolinguistic survey of Ghana's Central Volta Region. *Research Review* 12. 2–3.

- Saah, Kofi K. 1994. *Studies in Akan syntax, acquisition and sentence processing*. Ottawa: University of Ottawa dissertation.
- Sakel, Jeanette & Daniel L. Everett. 2012. *Linguistic fieldwork: A student guide*. Cambridge: Cambridge University Press.
- Schaefer, Paul A. 2009. *Narrative storyline marking in Safaliba*. Austin: University of Texas dissertation.
- Schuh, Russell G., 1995. Aspects of Avatime phonology. *Studies in African Linguistics* 24(1). 31–67
- Schaefer, Robert. 1975. *Collected field reports on the phonology of Frafra, No. 15*. Institute of African Studies, University of Ghana.
- Smye, Graham. 2004. *A grammar of Anufo*. (Language Monographs 7) Legon: Institute of African Studies, University of Ghana
- Snider, Keith L. 1989 *North Guang comparative wordlist: Chumburung, Krachi, Nawuri, Gichode, Gonja*. No. 4. Institute of African Studies,.
- Snider, Keith, & James Roberts. 2004. SIL comparative African wordlist (SILCAWL). *Journal of West African Languages* 31(2). 73–122
- Stewart, John M. 1966. *Awutu, Larteh, Nkonya, and Krachi with glosses in English*. (Comparative African wordlists no. 1) Legon: University of Ghana, Institute of African Studies.
- Thieberger, Nicholas (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford University Press.
- van Putten, Saskia. 2014. *Information structure in Avatime*. Nijmegen: Raboud University dissertation.
- Welmers, William E. 1946. A descriptive grammar of Fanti. *Language* 22(3). 3–78.
- Westermann, Diedrich H. 1907. *Grammatik der Ewesprache*. Berlin: Akademie Verlag
- Westermann, Diedrich H. 1922. Vier Sprachen aus Mitteltogo: Likpe, Bowili, Akpafu und Adele, nebst einigen Resten der Borosprachen. *Mitteilungen des Seminars für orientalische Sprachen* 25. 1–59.
- Westermann, Diedrich H. 1928. *Evefiala: Ewe-English Dictionary*. Berlin: Dietrich
- Westermann, Diedrich H. 1930. *A study of the Ewe language*. London: Oxford University Press
- Westermann, Diedrich H. 1954. *Wörterbuch der Ewe-Sprache*. Berlin: Akademie Verlag
- Zimmerman, Johann. 1858. *A grammatical sketch of the Akra or Gã language including vocabulary of the Akra or Gã language with an Adanme appendix*. Stuttgart: Basel Missionary Society.

Felix K. Ameka

f.k.ameka@hum.leidenuniv.nl

 orcid.org/0000-0001-9442-6675

Caucasus – the mountain of languages

Manana Tandashvili
University of Frankfurt

The widespread picture of linguistic diversity in the Caucasus as ‘the mountain of languages’ will be immediately confirmed if a closer look is taken at the region: multiethnic, multilingual, multireligious is the adequate description of this melting pot. What is responsible for the present-day ethnic, linguistic and sociocultural diversity is the historical coexistence of different ethnic groups in a geographically delimited region on the one hand, and the geopolitical situation at the border between the Orient and the Occident on the other. At the same time, this diversity leads to mutual influence of different kinds, ranging from linguistic and religious to ethnic assimilation. In this article, we will outline the results of relevant international projects in the field of ‘language documentation’ that we conducted over the past 15 years and what we have learned from these projects.

1. The study of the Caucasian Languages Research on Caucasian languages has been going on for quite some time. Even though Caucasian linguistics emerged a long time later than Indo-European, Semitic, or Uralian studies, it has a longer tradition than most other areas. Of the 70 odd languages that are spoken in the Caucasus, only Georgian can look back on centuries of uninterrupted written tradition, thus pertaining to the best documented languages worldwide, in contrast to other Caucasian languages whose written materials date only from the end of the Middle Ages (e.g., for Avar from the 14th, for Darginian from the 16th or for Tabasaran from the 17th century). All these materials mostly consist of word lists (compiled by scholars such as Güldenstädt, Klaproth, Peacock and others), which were often incorrectly transcribed because of a lack of knowledge of the Caucasian languages and are mostly unreliable as research materials.

Intensive studies of Caucasian languages began in the second half of the 19th century, commissioned by the Russian Academy of Sciences. Investigations by Peter von Uslar, Anton Schiefner, Adolf Dirr and others brought about numerous grammars of Caucasian languages such as Abkhaz, Chechen, Avar, Udi, and others. In addition to a description

of the language structure, these grammars often include concise dictionaries and texts of the respective languages.

In the Soviet era, the study of the Caucasian languages took a systematic and planned character: many languages, especially undocumented ones, were systematically recorded, fieldwork was conducted regularly, dictionaries were created and texts were published.

The Soviet stage of Caucasian language research was by all means a step forward in the documentation of Caucasian languages, but the scope of these efforts remained limited due to several factors such as the restricted personal competence of both speakers and scholars. The methods of recording the relevant materials were very differentiated and often inaccurate. Most of the time the people that were chosen as native speakers had a good competence of the language indeed, but the dynamics of the language situation was not considered, linguistic varieties being documented without any sociolinguistic background information (as to, e.g., language proficiency by age or social status of speakers).

In the 21st century, the development of technologies in language processing have completely changed the basis for documenting languages. In addition, theoretical approaches have been conceived that have facilitated new approaches, with Nikolaus Himmelmann's article of 1998 marking a decisive step forward, ensuring that Documentary Linguistics was institutionalized as a new subject of linguistic research, with a high impact on the Caucasian languages, too.

2. Endangered Caucasian languages in Georgia When the Volkswagen Foundation launched the call for proposals of the program DOBES ("Dokumentation bedrohter Sprachen" / "Documentation of endangered languages"), we developed a project at the Institute for Comparative Linguistics¹ of the University of Frankfurt, which aimed at identifying the endangered languages in one of the South Caucasian countries. We chose three languages spoken in Georgia to document and determine their degree of endangerment:

1. Udi, an East Caucasian language belonging to the Lezgi group, which is spoken in Georgia but also in Azerbaijan and elsewhere;
2. Batsbi or Tsova-Tush, an East Caucasian language belonging to the Nakh group, which is spoken only in Georgia, more precisely in just one village;
3. Svan, one of the South Caucasian languages, which is spoken mainly in Svanetia (in the Georgian high mountain area) but, because of recent migration, also in small groups (linguistic islands) in the lowlands of Eastern Georgia.

The DOBES project ECLinG (Endangered Caucasian Languages in Georgia), which dealt with the three aforementioned languages, was conducted in 2002-2006 in cooperation with Georgian partner institutions (the A. Chikobava Institute of Linguistics and the I. Javakhishvili State University, Tbilisi).

At the beginning of the project, when we started to record speakers of our object languages, we realised how strongly the minority language communities are characterised by multilingualism and what impact this has on the endangerment of their language. The research group fully shared N. Himmelmann's view that in the case of highly endangered languages, it is particularly important to document them, as further data collection in the

¹Since 2010 "Institut für Empirische Sprachwissenschaft" [Institute for Empirical Linguistics].

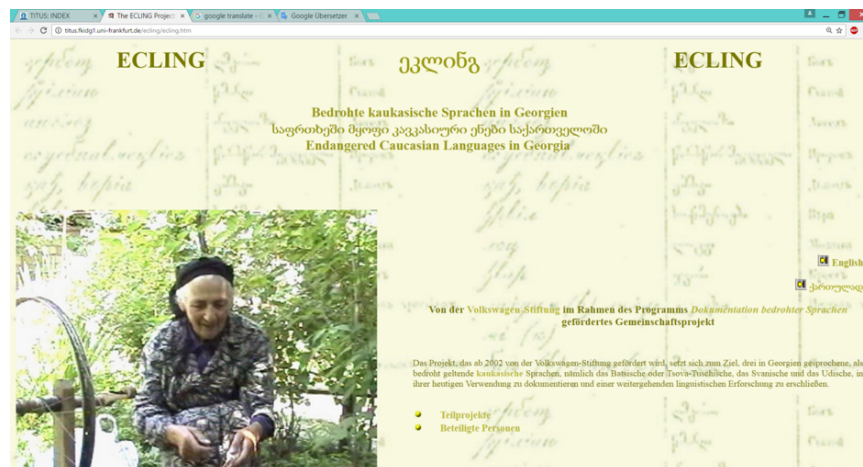


Figure 1: Screenshot of ECLing project website

future may be difficult if not impossible (Himmelman, Nikolaus P. 1998). For this reason, after the completion of the ECLInG project, we continued to document the small language communities in Georgia and continued our work in the new project “Sociolinguistic Situation in Contemporary Georgia” (SSGG; 2006-2009). The data collected in both projects are available in the TLA archive at the MPI for Psycholinguistics in Nijmegen.² In addition, the data have also been integrated into the databases of the TITUS project³ and the Georgian National Corpus (GNC). The results of our linguistic and sociolinguistic research have been discussed in various books and article (Chantladze et al. 2007–2010; Gippert 2008; Gippert 2012; Aristar-Dry et al. 2012; Tandashvili 2011; Tandashvili 2015).

For both scientific projects, our task was divided into two parts: language documentation and language description. In the following, I will discuss these two aspects using the example of two minorities in Georgia, the Udi and the Laz people.

3. Language description and (socio-) linguistic observations Within the ECLInG project, we were able to describe the language system of the three target languages and compile dictionaries of these languages. The text materials were first edited in Shoebox, later in Toolbox, with transcriptions, linguistic annotations, and translations into Georgian and English. In total, we have processed more than 70 hours of recordings in the ECLInG project.

On the basis of the audio-video recordings we hope to have prepared a solid base for research into all areas of the phonology, morphology, syntax etc. of these languages, but also their cultural properties. For example, we covered topics such as a) aspects of Christian and pre-Christian religion, b) material culture determined by a peculiar environment (the high-mountainous setting), with a focus on sheep breeding (in Batsbi), spinning, weaving and knitting (also in Batsbi), cattle breeding (in Svan), agriculture (in Svan and Udi), and wine growing (in Udi), but also c) folklore in form of singing and dancing (for Svan and Batsbi).

²The Language Archive. (<https://archive.mpi.nl>)

³*Thesaurus Indogermanischer Text- und Sprachmaterialien* (<http://titus.uni-frankfurt.de>)

While analyzing the recorded texts linguistically, we came across an interesting phenomenon that we call fluctuating or situational competence. This phenomenon is characterized by a different speech behaviour of a given speaker and shows different levels of competence depending on the situation: in a monologue the speaker cannot remember a certain word, changes the code and continues the talk in another (usually the dominant) language or uses foreign words from this language, while the corresponding lexical items are readily present in a natural (dialogic) speech situation—the word is ‘activated’ by this context and thus used without hesitation.

Often, for example, Udi speakers are not familiar with specific agricultural terms. In this case they use Georgian lexemes, integrating them in accordance with the verbal morphology of Udi, but at the same time they are aware of this process and sometimes explicitly signal it as being a change of code:

- (1) *Oša me čapnux gapurčna-yan-besa...*
 ‘Then we thin this vineyard out...’ (in Udi)
es ukve kartuli siťvaa.
 ‘this is already a Georgian word’ (in Georgian)

It is further important that code-switching is a natural thing in the multilingual environment of the Udis, often provoked by the incompetence of the speaker:

- (2) *Tene hebsa, davateyanduxsa, ič benginänne, täksä heba... gasxvla rogor aris beš muzin, tezaaba.*
 ‘Nothing is necessary, we do not disinfect, it grows by itself, alone ... What *gasxvla* (trim the vines) is in our language, I don’t know.’

However, as the next example shows, this ‘incompetence’ can be overcome if the speaker uses the word in a natural context (here during the vintage in October):

- (3) *Siftä kaçyanexa čapnu, toxyanbesa, davyanduxsa.*
 ‘First we cut the vine, we weed it, we disinfect it.’

Linguistic behavior in extra-contextual cases is often accompanied by code switching. Based on our materials, we have divided the manifestations of code switching into two types:

1. ‘signalled’ code switching, i.e. code switching that is caused by triggering words and is often associated with incompetence of the speakers;
 2. ‘non-signalled’ code switching, which is exerted between two languages without a signal and without damaging the ecology of the language. The switching of the code can e.g. be carried out in the middle of a subordinate clause as in example (4); here the boundary between the two languages and the syntactic boundaries in the sentence do not coincide.
- (4) ... { [*gasxvla rogor aris*]_{Georgian} [*beš muzin, }_{SUB} tezaaba*]_{Udi}
 [what ‘trimming’ is]_{Georgian} [in our language, I don’t know]_{Udi}

With Udi speakers, code switching can be observed not only within sentences but also within numerical expressions. In our recordings we have documented cases where ones and tens are composed using components from different languages, as in *qər biḫ* ‘forty-four’, composed of Azeri *qər* and Udi *biḫ*.

A particularly interesting issue of the project was to compare the language situation of the two ‘insular’ varieties of Svan in the Kodori valley and in the lowlands. While Kodori Svan does not show any specific changes in the grammatical system or in the lexicon, the lowland variety seems to be far more endangered, although in both cases the children attend Georgian schools and their education as well as the available media use Georgian only. This indicates that the geographic seclusion of a linguistic island with little contact with the main language area is not necessarily the primary cause of increasing endangerment. Rather, the variety in question can evolve with its own language structure and, in extreme cases, end as an independent dialect. However, an island whose sociolinguistic context is determined by a dominant surrounding language can show within two generations all the characteristics of endangerment, which manifest themselves in the abolition of the continuous transmission of language and cultural heritage.

Among the Svan people now living in the eastern Georgian lowlands, we have seen yet another phenomenon, namely a special case of multilingualism, which was also observed later on in the Laz community in the framework of the SSGG project. In the conversation between the older and younger generation often two languages are used: the grandfather speaks with the grandchild in Svan or Laz; the child understands what has been said but cannot answer in the same language and so uses Georgian. Thus the entire dialogue takes place in two languages, and each participant can be assigned one of them.

Comparing the linguistic situation of the Batsbi people in Georgia with that of the Udi and Svan people, one finds another case of assimilation, which will be discussed in the following section.

4. Collecting data for multidisciplinary research As the examples presented above show, the data collected in our project can serve not only as the basis for (socio)linguistic studies but also provide much material for other disciplines such as ethnology, ethnopsychology, ethnography, oral literature, religious history, conflict research, etc.

In order to describe the language situation more accurately and to document the Udi language in its fields of use, we tried to observe and continuously document Udi speakers for more than 10 years, visiting the locations at least three times a year—in spring (March–April), in summer (July–August) and in autumn (in October).

While the first recordings were still relatively spontaneous and unplanned, aiming to become acquainted with important topics such as personal biographies, household, everyday activities or fields of employment and to collect the relevant vocabulary, in later years we took recordings with more specific goals. So, we filmed Udi people e.g. during the vintage or went to a wedding in order to document the natural process of this ritual. In 2011, we attended the inauguration of the first church in Zinobiani asking the Udi people about the importance of religion with regard to their (cultural) identity.

All these recordings show a remarkable asymmetry between the various components of identity (linguistic, ethnic, and religious), which I studied in my essay “What causes the endangerment of languages? The Case of the Udi Language in Georgia” (Tandashvili, Manana, 2018). The exploration of the language situation of the Udi and the Batsbi

people has clearly shown that the linguistic identity does not necessarily have to be linked to an ethnic identity, just as the ethnic identity does not necessarily derive from a linguistic identity. Thus, the Udi people are almost completely assimilated linguistically, but ethnically they have maintained their identity. The Batsbi people, conversely, are assimilated ethnically by identifying themselves as Georgians but have kept their identity linguistically. This results in a particular identity constellation: the terms 'Batsbi people' and 'Svan people' cannot be used as ethnic terms but refer to the linguistic identity of the community concerned, while 'Udi people' in Georgia characterizes an ethnic group.

Assuming that the ethnic identity is based on 'historical memories', the 'ethnic stability' of the Udi people can well be understood. However, it is also known that this type of collective memory can be forgotten, revitalized, or reshaped. The 'historical memory' of the Batsbi people is largely erased, while the Udi people have revitalized theirs by referring to the medieval state of 'Albania' in the Caucasus. The discovery of the Caucasian Albanian manuscripts in St. Catherine's Monastery on Sinai, on the one hand, and the idea of being descendants of one of the oldest people in the southern Caucasus, on the other, today plays a major role in the cultural identity of the Udi people. Whether the official recognition of their ethnicity has helped here remains unanswered and requires further investigation. About the Batsbi people, however, it can be said by now that their 'historical memory' is deleted.

References

- Aristar-Dry, Helen, Sebastian Drude, Menzo Windhouwer, Jost Gippert & Irana Nevskaya. 2012. Rendering Endangered Lexicons Interoperable through Standards Harmonization: the RELISH Project. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 766–770. Istanbul: European Language Resources Association.
- Chantladze, Iza, Ketevan Margiani-Dadvani, Ketevan Margiani-Subari, Medea Sagliani & Rusudan Ioseliani. 2007–2010. *The Kodori Chronicles. (With Researches). Vol. I*. Tbilisi.
- Gippert, Jost. 2008. Endangered Caucasian Languages in Georgia. In David K. Harrison, David S. Rood & Arienne Dwyer (eds.), *Lessons from Documented Endangered Languages*, 159–194. (Typological Studies in Languages 78). Amsterdam: John Benjamins.
- Gippert, Jost. 2012. Language-specific encoding in endangered language corpora. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Potentials of Language Documentation: Methods, Analyses, and Utilization*, 17–24. (Language Documentation & Conservation Special Publication, 3). Honolulu: University of Hawai'i Press.
- Himmelmann, Nikolaus P. 1998. Documentary and Descriptive Linguistics. *Linguistics* 36. 161–195.
- Tandashvili, Manana. 2009: Georgien – multiethnisch, multilingual, multireligiös. *PRO GEORGIA. Journal of Kartvelological Studies* 18. 51–65.
- Tandashvili, Manana. 2011. Change of Codes under mulilingual conditions and idiolectal Interferences. In Iza Chantladze, Eter Soselia, Nugsar Mgeladze (eds.), *Linguocultural Researches*, 161–171. Batumi: Shota Rustaveli State University Press.
- Tandashvili, Manana. 2015: Sprachliche Minderheiten in Georgien (Zur Methodologie). In Natia Reineck & Ute Rieger (eds.), *Kaukasiologie heute. Festschrift für Heinz Fähnrich zum 70. Geburtstag*, 395–410. Jena.
- Tandashvili, Manana. 2018. What causes the endangerment of languages? The Case of the Udi Language in Georgia. *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 15. 289–312.

Digital Resources

The Language Archive

(<https://archive.mpi.nl>)

Thesaurus Indogermanischer Text- und Sprachmaterialien

(<http://titus.uni-frankfurt.de>)

ECLinG - Endangered Caucasian Languages in Georgia

(<http://titus.fkidg1.uni-frankfurt.de/ecling/ecling.htm>)

SSGG - Die soziolinguistische Situation im gegenwärtigen Georgien

(<http://titus.uni-frankfurt.de/ssgg/ssgg.htm>)

BaLDAR Batumi Linguocultural Digital Archive

(<http://digiarchive.bsu.edu.ge/>)

Manana Tandashvili
tandaschwili@em.uni-frankfurt.de

Reflections on language documentation in India

Shobhana Chelliah
University of North Texas

The last twenty years have seen efforts to support the study of minority and lesser-studied languages of India from varied stakeholders: these include the Indian government, international and Indian nonprofit organizations, indigenous and state-level cultural and language committees and institutes, and individuals with a passion to preserve and document their cultures and languages. Their efforts have led to mixed success due to conflicting ideologies, history, and resource availability (Annamalai 2003). Basing my observations on my research, personal experience and engagement with language documentation activities in the country, I provide an overview of the current state of language study and my hopes and efforts for future of language documentation and description in India.

1. A philosophy of language science and its consequences Woodbury defines language documentation as “the creation, annotation, preservation, and dissemination of transparent records of a language” (2011:159). While there is plenty of traditional data gathering toward language description in India (Abbi 2001), the other facets of language documentation mentioned in Woodbury’s definition are still emergent.

A common viewpoint in linguistics departments in India is that extended and varied data collection, such as that needed for the creation of a documentary corpus, is time consuming and not useful for scientific publications. As result, much of the language data collected is at the level of the word or clause and is collected through responses to questionnaires. Language science, produced and sanctioned in this way, is reflected in the very successful society and related summer school, known as the Formal Studies in the Syntax and Semantics of Indian Languages (FOSSSIL). The stated goal of FOSSSIL is to:

undertake schemes/activities leading to the development of formal linguistic descriptions and analyses of Indian languages and to provide facilities and act as a forum for exchange of information, ideas and experience in the

practices and techniques in formal linguistic descriptions and analyses of Indian languages.¹

Many of the leading linguists in India are on the governing body of FOSSSIL and likewise, many of the programs around the country have this same focus. A consequence of this focus is the lack of encouragement towards resource creation, especially the lack of language data in the form of annotated corpora. Annotated corpora, however, are central to the goals of language documentation since such corpora provide the interpretive apparatus to audio and video records. When the corpus includes a variety of genres and events, even without audio and video, the corpus can illustrate and linguistic and cultural practices for future generations.

In the Indian context, annotated corpora could provide data sources to transform language science in the region. Looking at the descriptive theses and dissertations produced over the last 20 years, we see that linguists-in-training need more scaffolding in their attempts at language description. With only translation to guide data gathering, these researchers often miss less observable or reportable grammatical features like evidential systems (I discussed this in some detail in Chelliah 2001). Naturally occurring language samples could stimulate new avenues of grammatical investigation and thus prompt new linguistic discoveries. An easy way to do this would be to require Ph.D. theses that use minority or lesser-studied language data for typological, descriptive, or other analytic argumentation to include annotated texts in appendices with data cross-referenced in the body of the thesis. Another possibility would be to establish a new format for MA theses where these are not just descriptive sketches but are in the main annotated corpora, collected, analyzed, and annotated using the gold standard for this type of data analysis and including a grammatical sketch (see Woodbury 2014). This would serve several purposes: it would challenge authors to move beyond the fill-in-the-blanks method of grammar writing to one that attempts to account for patterns found in the corpora; provide training in transcription, annotation, analysis, and translation; and create useful documentary materials.

2. National infrastructure and linguistic culture India, through policy and practice, has privileged its larger languages, so it is unclear how efforts at documenting minority and lesser-studied languages will overlay with this linguistic culture in the long term. Indian states were reorganized based on language after independence (King 1997). National policy is in place to recognize languages and provide resources for their maintenance if the language: (1) Has an ancient literary tradition, in which case it falls under the “classical” language category; (2) Has speaker population of over a million and a writing system, which can support inclusion in the constitution’s schedule of national languages (the 8th schedule). Twenty-two languages have this status and afford speakers rights to early childhood education and the ability to take national exams in that language; (3) Has a “pure” variety, which is often the higher register in a diglossic situation. The effort to preserve such purity is the ideology that led to the curation and preservation of ancient Vedic hymns in the first millennium and led to the need for the speaker of other varieties to have the interpretative grammatical insights offered by Panni (Scharf 2013, Schiffman 1996:152). In an environment increasingly hostile to the secular ideals of the Indian constitution, ideologies that use Hindi as a symbol for Hinduism have contributed to an imagined dialect continuum that has impeded accurate language identification. For

¹http://www.fosssil.in/index_fosssil.htm

example, the 2001 Indian Census assigns membership of 47 varieties as dialects of Hindi and overall 1652 varieties into 122 languages (Abbi 2004: 2).

On the other hand, the Indian government both indirectly and directly supports language documentation of minority languages. An example of indirect support that affirms language documentation as a worthy activity was seen in 2013 when the Indian government conferred a high national honor, the Padma Shri, to Anvita Abbi for her work documenting the languages of the Andaman.

Examples of direct support for language documentation and dissemination are through infrastructure such as the creation of institutes and projects. The Sahitya Akademi, an organization dedicated to the promotion of the literature of major Indian languages, recently added a Centre for Oral and Tribal Literature. Under the direction of Anvita Abbi in 2015–2017, this center released several publications on lesser-known languages, initiated a new series titled *Unwritten Languages*, and started a digital collection of oral literature. The Indira Gandhi National Centre for Arts Cultural Archive (IGNCA) located in New Delhi houses artifacts of anthropological interest including language materials such as microfiche of pre-20th century manuscripts from North East India. The IGNCA includes a growing digital repository.

The Office of the Registrar General and Census Commissioner under their Language Division oversees the Linguistic Survey of India (LSI). Initiated in 1981, this work is an ambitious updating of the pre-independence Linguistics Survey of India compiled, edited, and published by George Grierson between 1903–1928. The volumes produced by LSI are extensive in the coverage of demographic information, list of languages and numbers of speakers including a type of social network analysis laying out which languages are home languages, and which languages have wider function. Also included are maps and sociolinguistic information. The linguistic descriptions include useful overviews of nominal and verbal morphology including verb conjugation and simple sentence types. Descriptions are more detailed for larger languages like Oriya (42 pages), than for smaller languages like Relli (23 pages). The format for all are the same since they were collected using a standardized survey: a word list, sentence list, and a story in English for the connected text sample which appears to be the Prodigal Son from the book of Matthew in the New Testament Bible. This is one of the texts translated in the Grierson LSI.² The preface of one volume also notes that, “the whole of the interview was recorded us[ing] a tape recorde[r], the recording of which were transcribed in the field, using narrow phonetic transcription based on the International Phonetic [Alphabet] (Banthia 2002: viii)”. Although some of the original Grierson LSI gramophone recordings are available online through the University of Chicago South Asia Digital Library, the modern LSI recordings are not publically accessible in either analog or digital format.

In 2007 Government of India also earmarked 280 crore (approximately \$40 million USD in 2018 conversion rates) towards the documentation of endangered languages. The Central Institute of Indian Languages (CIIL) in Mysore was commissioned to oversee the training of hundreds of field linguists to document speech varieties village-by-village to gather information on language structure, script, and literature (Srivatsa 2012). This project is no longer in place as it was first conceived although work on endangered languages continues at CIIL. In 2013, the Ministry of Human Resource Development initiated the Scheme for Protection and Preservation of Endangered Languages (SPPEL) with the immediate goal of providing a grammar, dictionary, and ethnolinguistic sketch

²PDFs of these descriptions and the preface information are viewable Linguistic Survey of India website: www.censusindia.gov.in/2011-documents/lsi/ling_dnh.html

for 117 languages of 10,000 or fewer speakers with a long-term goal of covering 500 languages.³ CIIL, in collaboration with academic and cultural institutions, is leading the implementation of this scheme across India. While the SPPEL website includes ethnolinguistic notes; sample audio clips from word lists with transcription; and a bibliography for some of the listed languages, progress towards the goal of grammars and dictionaries appears to be slow. SPPEL has also undertaken a massive effort to train native speakers and younger linguists in data elicitation, audio recording, and acoustic analysis. A related annual conference, The International Conference on Endangered and Lesser Known Languages, features invited lectures by national and international documentation and archiving experts.

The University Grants Commission in the 2000s set up centers to support endangered languages: for example, the Centre for Endangered Languages of Northeast in Tezpur University in Assam, and the Centre for Endangered Languages & Mother Tongue Studies at the University of Hyderabad in Telangana. In addition to documentation and preservation, the vision for the centers is to instill in speakers an appreciation for their languages and to support language transmission through creation of pedagogical material. This linking of social responsibility and linguistic work is relatively new for India and is articulated with increasing frequency. An example is the purpose statement of the new conference, Approaches and Methodologies for the Study of Indigenous and Endangered Languages, which states:

Language is an integral part of the social identity and ethnicity of a group. In order to preserve a social group's or tribe's cultural identity, it is essential to understand the urgency and draw a roadmap for preserving their cultural, social and linguistic heritage and identity.⁴

To these newer dedicated centers, we can add established programs that have been offering field-based linguistics courses: Jawaharlal Nehru University and Delhi University in Delhi; the Central Institute of Indian Languages in Mysore; the University of Guwahati and Tezpur University in the Assam; and Chandigarh University in Punjab.⁵

Language documentation in India is taking place where two contesting ideologies exist: one that articulates the value of linguistic diversity and the other that supports the large and religiously relevant. Speakers of minority languages contest dominant ideologies through grassroots movements valorizing their languages and affirming cultural and group identity through language. At the same time, some of the same groups support the dominant ideology by, for example, lobbying for and gaining acceptance to the 8th Schedule of languages (e.g., Meiteiron (Manipuri) with approximately 1.2 million speakers was included in the 8th schedule in 2005). The fluid alignment and misalignment with state policy is reflected in the ways we affirm identity in India, from code switching, code selection, conformity and resistance.⁶

³www.sppel.org/soligadoc.aspx

⁴<https://linguistlist.org/issues/29/29-414.html>

⁵The existing literature on the ethical conduct in language documentation still focuses on the “white researcher and native community”. The Indian context needs consideration of a much more complex mix of “researchers”, including missionaries, literacy experts, national surveyors, national and international academics, speaker-academics, and citizen-documenters.

⁶I discuss this in reference to name choice in Chelliah (2005).

3. Institutional incubators and grassroots movements In parallel with formal linguistics programs in India, language documentation and preservation activities are numerous.

At the Northeast Indian Linguistics Conference which meets in Assam and nearby states every two years, I learned that many communities are eagerly pursuing orthography development and dictionary creation, some with the help of missionary groups interested in Bible translation. I have written about individuals who have worked to document their cultural events and practices with minimal training (Chelliah 2016). In the Lamkang community in Manipur state, Mr. Beshot Khular has produced three books and a set of audio and video DVD on proverbs, traditional narratives, and traditional song and dance. Reverend Daniel Tholung recorded elders telling traditional stories and created film on dances during major festivals to capture specific outfits, headdresses, and other ornamentation used during those dances. He paid close attention to the vocabulary used during these events and to the specific uses of objects. Another remarkable documenter is Mr. Somi Roy, the founder and managing trustee of the Imasi Foundation. The mission of the IMASI foundation is to promote Manipuri culture through documentation, preservation, and dissemination of literature. Recently, Mr. Roy, an accomplished translator of literary works and screenplays from Manipuri to English, has added verbal art to his list of interests. For example, he is creating online resources on a cycle of songs sung in classical Manipuri Sankirtan style by an all-women's choir called the Jalakeli performance. His work fits well with Woodbury's definition of documentation: videos of performance of the songs of the Jalakeli; interviews with the singers of the Jalakeli; and documents about the Jalakeli tradition. At Guwahati University, we find under the direction of Professor Jyotiprakash Tamuli, a growing cadre of students working with both Indian and non-Indian linguists to produce substantial documentation such as Krishna Boro's descriptive grammar of Hakhun Tangsa and Prafulla Basumatary work on Boro grammar, both with accompanying audio and video documentation. There are many others inspired by the atmosphere and training provided at Guwahati University, who work with their communities to document their language.

Two high-profile ventures to "document" language and culture in India have brought India's linguistic diversity to the public's attention. Ganesh Devy, once a professor of English and now self-taught linguist, is the conceptual and practical lead for the People's Linguistic Survey of India (PLSI). Devy, organizing under a nonprofit called the Bhasha Research and Publication Centre, utilized the effort of 3500 volunteers including native speakers or speakers of related languages, linguists, and historians to gather language information state-by-state. The group has identified 780 languages and plans to publish collections of descriptions to cover all the languages they have identified. Since there is no accompanying audio and video recording along with the publications, the activities for PLSI do not fit the definition of language documentation laid out in Himmelmann's seminal article to which this volume is dedicated. However, we can recognize PLSI for starting a popular conversation in India on linguistic diversity, prompting the former Indian Prime Minister Manmohan Singh, to state in a public lecture that India is ignoring its linguistic diversity and that more should be done to bring minority languages to the digital age (Banerjee 2017).

An International nongovernmental organization, the Living Tongues, has also been active, creating dictionaries with audio accompaniment on Munda and other languages

of Arunachal Pradesh.⁷ Living Tongues provides workshops to train speakers on how to add to existing lists of words so that both audio and transcription can be crowd sourced. The long-term effects of this training are yet to be seen - it is hopeful that these interventions will create local documentary linguists and stimulate language activism that will result in long lasting useful language documentation. I should also mention that there are many individuals from universities worldwide that work on documentation projects in India. The major traffic for PhD documentation work is from Singapore (Nanyang Technological Institute); Thailand (Payap); Australia (Melbourne and Sydney); the UK (the School of African and Oriental Studies); Switzerland (University of Zurich); Germany (Cologne and Max Planck, Nijmegen); and many universities in the United States. Stephen Morey (Melbourne) and Mark Post (Sydney) have provided training for local linguists on a consistent basis for several years and the quality of curation practices has greatly improved due to their efforts. Missionary groups from the United States, especially those related Summer Institute of Linguistics are in many parts of India doing literacy work which often includes dictionary creation and orthography development.

4. Building on strong beginnings We can see that there are enormous resources available to engender lasting and valuable language documentation in India. In my view, all that is needed now to spring forward from traditional data gathering for language description to data gathering for language documentation is a culture that rewards archiving of primary source materials. We need to shift the focus from training that results in quick-fix grammatical sketches to training that results in rich documentation that we can later harvest for description for pedagogy, revitalization, or science. That is, we need to focus on audio, and video data collection, metadata creation, data management, and data curation. A resource for training, discussion, and sharing on language documentation, such as the US Institute on Collaborative Language Research⁸ or Summer School in Language Documentation and Linguistic Diversity⁹ specially tailored for India would be hugely supportive of this growth.

Another need are local repositories and national or at least regional archives for digital language data. I am inspired by the many native speaker-linguists and language activists who work, in spite of all manner of obstacles, to write, read, publish, teach, proclaim, and celebrate their languages and cultures. At the University of North Texas (UNT), I am working with colleagues to create a repository for annotated corpora of South Asian languages which we are calling the Computational Resource for South Asian Languages (CoRSAL). With CoRSAL, my hope is to establish the value of annotated corpora (with source audio and video) for multiple fields of study and for multiple stakeholders and to raise the prestige of creating such corpora. We plan for the CoRSAL repository to house data in formats that can be easily accessed and used by indigenous groups, linguists, and other researchers for a range of purposes, including language science, computational linguistics, language reclamation and revitalization, language teaching, and investigations into diverse cultures and histories. We banking on the idea that “If we build it, they will come”! The CoRSAL concept has already spun off into interesting subprojects led by my UNT colleagues: metadata improvement for language archiving by Oksana Zavalina; user-centered design of language archives by Christina Wasson;

⁷<https://livingtongues.org/india2015/>

⁸<https://en.wikipedia.org/wiki/CoLang>

⁹<http://www.bu.edu/applied-linguistics/2014/01/23/summer-schools-international-summer-school-in-language-documentation-and-linguistic-diversity-stockholm-sweden/>

information seeking behaviors of indigenous populations by Mary Burke; and shared data formats for cross corpora comparison by Alexis Palmer, Manish Srivastava (from the International Institute of Information Technology – Hyderabad), and me. In addition, many of my colleagues working in Northeast India, senior and junior, are partnering in the project by contributing existing corpora from languages spanning from Arunachal Pradesh to Tripura. We have partners in Assam with whom we are working to create complimentary archives for audio and video and accompanying annotated materials. We hope to find funding not only for the CoRSAL concept but for grants to train and support documentation.

True collaborative language documentation is still a rarity in India. But my belief is that progress in language science and documentary linguistics is going to come from communities seeking support from linguists to create resources for revitalization. These agents of change will take us out of the doldrums and blow us into a storm of linguistic discovery. I hope to be right in the middle of that storm.

References

- Abbi, Anvita. 2004. Vanishing diversity and submerging identities: An Indian case. Paper presented at the *conference of Dialogue on Language Diversity, Sustainability and Peace*, 20–23, May 2004. Barcelona, Spain.
- Abbi, Anvita. 2001. *A manual of linguistic fieldwork and Indian language structures*. Munich: Lincom Europa.
- Annamalai, E. 2003. The opportunity and challenge of language documentation in India. In Peter K. Austin (ed.), *Language Documentation and Description Volume 1*, 159–167. London: School of Oriental African Studies.
- Banerjee, Rumu. 2017. Manmohan Singh stresses on the need to tap potential of India's linguistic diversity. *The Times of India*, August 4 2017. (<http://timesofindia.indiatimes.com>)
- Banthia, Jayant Kumar. 2002. Foreword. In S. P. Datta (ed.), *Linguistic survey of India special studies Orissa, p. 5*. Kolkata: Language Division, Office of the Registrar General.
- Chelliah, Shobhana. 2005. Asserting nationhood through personal name choice: The case of the Meithei of Northeast India. *Anthropological Linguistics*, 47(2). 169–216.
- Chelliah, Shobhana. 2001. The role of text collection and elicitation in linguistic fieldwork. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 152–165. Cambridge: Cambridge University Press.
- Chelliah, Shobhana. 2016. Responsive methodology: Perspectives on data gathering and language documentation in India. *Journal of South Asian Languages and Linguistics* 3(2). 176–196.
- Grierson, George Abraham (ed.). 1903–1928. *Linguistic survey of India*. Calcutta: Office of the Superintendent of Government Printing.
- King, Robert D. 1997. *Nehru and the language politics of India*. Delhi: Oxford University Press.
- Scharf, Peter M. 2013. Linguistics in India. In Keith Allan (ed.), *The Oxford handbook of the history of linguistics*, 227–258. Oxford: Oxford University Press.
- Schiffman, Harold. 1996. *Linguistic culture and language policy*. London: Routledge.
- Srivatsa, Sharath S. 2012. New Linguistic Survey of India to begin in April next year. *The Hindu*, Updated: March 22, 2012 10:27. (<https://www.thehindu.com/>)
- Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.
- Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan & Peter K. Austin (eds.), *Language documentation and description, Volume 12: Special Issue on Language Documentation and Archiving*, 9–36. London: School of Oriental African Studies.

Shobhana Chelliah
shobhana.chelliah@unt.edu

Reflections on linguistic fieldwork and language documentation in eastern Indonesia

Yusuf Sawaki

Center for Endangered Languages Documentation, University of Papua

I Wayan Arka

Australia National University

Udayana University

In this paper, we reflect on linguistic fieldwork and language documentation activities in Eastern Indonesia. We first present the rich linguistic and biological diversity of this region, which is of significant interest in typological and theoretical linguistics and language documentation. We then discuss certain central educational issues in relation to human resources, infrastructures, and institutional support, critical for high quality research and documentation. We argue that the issues are multidimensional and complex across all levels, posing sociocultural challenges in capacity-building programs. Finally, we reflect on the significance of the participation of local fieldworkers and communities and their contextual training.

1. Introduction In this paper, we reflect on linguistic fieldwork and language documentation in Eastern Indonesia. By “Eastern Indonesia,” we mean the region that stretches from Nusa Tenggara to Papua,¹ including Nusa Tenggara Timur, Sulawesi, and Maluku. This region is linguistically one of the most diverse regions in the world in terms of the number of unrelated languages and their structural properties, further discussed in the next section. This is the region where Nikolaus Himmelmann has done his linguistic

¹The term “Papua” is potentially confusing because it is used in two senses. In its broad sense, it refers to Indonesian Papua, formerly called Irian Jaya. Indonesian Papua has now been split into two provincial units, Papua Barat and Papua, with Papua Barat covering the western areas of Indonesian Papua, from the Teluk Wondama and the Kaimana regencies in the southeast to the Raja Ampat regency in the west. It covers the entire Bird’s Head region of New Guinea. In its narrow sense, Papua refers to the eastern part of the former Irian Jaya.

fieldwork and language documentation, investing his tremendous efforts in this endeavor for about thirty years (1980s–2018). Especially in Papua, he has worked with local linguists, linguistic students, and language community members through the Center for Endangered Languages Documentation (CELD), which he established at the University of Papua with the first author of this paper. Our reflections in this paper are slightly biased by our own linguistic fieldwork and modern language documentation in Papua (Barat) and Nusa Tenggara. Both of us are Indonesians, fortunate to have received high-quality training overseas. We can therefore reflect on linguistics and language documentation in Eastern Indonesia from two perspectives, as insiders and outsiders. We begin by considering the rich linguistic diversity of Eastern Indonesia, followed by our discussion of the issues faced in conducting fieldwork and language documentation in this region with respect to local collaboration, capacity building, and institutional support.

2. Eastern Indonesia: An area of rich biodiversity and ethnolinguistic diversity

The region of Eastern Indonesia, often called East Nusantara,² is home to about 500 of the country's approximately 700 languages; genealogically and structurally diverse (Arka 2016; Holton and Klamer 2017), they serve as living laboratories for linguistic research.

Genealogically, the languages in this region belong to the two major families—Austronesian and non-Austronesian (or Papuan)³—with their precise subgroups still needing further research (cf. Blust 2009; Donohue and Grimes 2008). In certain areas, particularly North Halmahera, Timor-Alor-Pantar, and Bird's Head, the languages of these two major families have been in contact over millennia. The different waves of migration of both Austronesian and Papuan peoples in this region are evidenced by archeological and linguistic data (Bellwood 1997, 123; Klamer and Ewing 2010, 3; Ross 2005, 42). Over millennia, these languages have undergone gradual diversifications, resulting from extensive dialectal variations in so-called dialect chains, forming a linkage (cf. Ross 1988, 9–11) but with no discrete proto-language, hence the debate on the existence of Central Malayo-Polynesian, for example (Blust 1993, 2009; Donohue and Grimes 2008; Klamer 2002a, 2002b). The linguistic complexity in Eastern Indonesia, which has resulted from a combination of contact-induced changes and other kinds of internal diversifications, has posed a challenge and will remain so in historical linguistics for years to come.

The diverse structural properties of the languages in this region are of primary interest in typological and theoretical linguistics. Some languages are highly isolating, typically in Flores, such as Rongga (Arka 2015) and Keo (Baird 2002); others are morphologically complex and fusional, such as Marind (Ndiken 2011; Olsson 2017), and several are agglutinative, such as Wooi (Sawaki 2016). Salient features of Austronesian and Papuan languages in this region (for details, see Hajek 2010; Himmelmann 2005; Holton and Klamer 2017; Klamer 2002b) include relatively simple phonemic inventories, commonly with five basic vowels and various simple consonantal systems with an average inventory size of sixteen (Hajek 2010, 27–28); the presence of implosives and/or prenasalized

²East Nusantara refers to the geographical region to the east of the Wallace Line, covering the areas of Sumbawa, Sulawesi, Nusa Tenggara, Maluku, including Halmahera, and Bird's Head of New Guinea, as well as East Timor (Klamer and Ewing 2010: 1; Klamer and Kratochvil 2014: 1).

³As a linguistic term, Papuan has generally characterized all groups of unrelated languages that do not resemble Austronesian and Australian. Papuan languages stretch from Alor-Pantar to Halmahera in the west to the mainland of New Guinea and adjacent islands to the east of the mainland of New Guinea (Bellwood 1995; Foley 2000; Spriggs 1997; Wurm 1982). Linguistically, they feature diverse, structurally complex, and genetically unrelated languages but commonly share certain typological and structural similarities, setting them apart from the languages of different groups in this area.

consonants; and quite complex morphology, phonology, and morphosyntax (particularly in Papuan languages), with argument indexing, verbal serialization, and the absence of a grammatical subject/pivot. The diversity of these linguistic features expressed in the different languages in this region is of immense interest and significance in the field of linguistics. Many of the languages in this region are still under-documented or undocumented. Further documentation is needed and particularly urgent for the highly endangered languages. New data from these languages will add to the empirical basis required in linguistic typology and theoretical linguistics in general and promise a breakthrough in the understanding of the history of two groups of languages (the Austronesian and the Papuan languages) the people in the region, and the extent of variability in human language.

Eastern Indonesia and West Timor also constitute a region of mega biodiversity, which correlates with ethnolinguistic diversity. It covers the area of Wallacea, bioecologically known as a transition between Sundaland (the Malay Peninsula, Sumatra, Borneo, Java, and Bali) and Sahul (Australia and New Guinea). It is defined by its rich mix of biodiversity—flora and fauna—with some Asian, Australian, and several unique endemic types. Ethnographically, the areas of Wallacea and New Guinea are home to diverse ethnic groups, showing rich Austronesian and Melanesian cultures, often with a mixture of the two cultures due to internal diversifications and contacts among the Austronesian and the Melanesian people. Such contacts that result in inculturation are observed in Alor-Pantar (Klamer 2008), Halmahera (Bowden 2001; Platenkamp 1990), and Bird's Head of New Guinea (Klamer 2002a; Timmer 2002), for example.

Some researchers affirm the correlation between biological and ethnolinguistic diversity (Harmon 1996; Harmon and Maffi 2002; Turvey and Pettorelli 2014). Schapper (2015) shows that while the rich biodiversity of Wallacea correlates with its rich linguistic diversity, the region also constitutes a linguistic area (i.e., with shared sets of features). Language plays a central role in the transmission of the local knowledge related to this biodiversity and ethnolinguistic diversity. However, current unprecedented changes in the physical ecology of Eastern Indonesia (e.g., the end of isolation with the construction of new roads, accompanied by an influx of immigrants) and related sociopolitical developments in modern Indonesia have threatened this region's biodiversity and ethnolinguistic diversity. In this regard, language documentation is a matter of urgency. Our ethnobiological documentation projects funded by the Endangered Language Documentation Program (ELDP) are part of the efforts to record ethnobiological and ecological data. For example, we have conducted documentation research on the ethnobotanical, economic, and linguistic aspects of *sago* (*Metroxylon sago*) in Marori (Hisa, Mahuze, and Arka 2017) and mangroves in Wooi (Sawaki 2016). Local folklore and stories also contain rich information about how local communities traditionally live and manage their physical environments, for example, narratives about the Emayode clan, *sago* in Kokoda, and the south coast of Bird's Head of New Guinea (Sawaki 2017). All these are in line with the language documentation principle of recording "as many and as varied records as practically feasible, covering all aspects of the set of interrelated phenomena commonly called a language" (Himmelman 2006, 2).

3. Local linguists and leadership Reflecting on the role of local linguists and project leaders in language documentation in Eastern Indonesia, we need to examine its history and current situation. Historically, linguistic fieldwork and language documentation with a recognized impact have been led and carried out by foreigners. The traditional linguistic

work in Indonesia was started in the late nineteenth century and continued to the early twentieth century by Christian missionary linguists, joined by non-missionaries, mainly academics (university-based and independent linguists) affiliated with various foreign institutions. Through their work, various publications have been produced, with subjects ranging from comprehensive grammar to topic-related descriptions. Over the last twenty years (from the late 1990s to 2018), the areas of Papua and Nusa Tenggara have been under intensive study, with linguists focusing on modern language documentation, including documentation projects undertaken by the first author for the Woi language in Papua (2009–2012) and by the second author for the Rongga language in Flores (2004–2006) and the Marori language in Papua (2016–2017). The projects have produced a number of multimedia files (deposited in the Language Archive⁴ and ELAR⁵) literacy materials for local communities, bilingual dictionaries, and grammar books.

While some traditional work has been carried out by *Badan Bahasa* (National Language Board), most modern language documentation in Eastern Indonesia has been undertaken by foreigners. Ideally, language documentation should be led and performed by Indonesian linguists from the community or at least by non-local Indonesian linguists, for at least two reasons. First, employing Indonesian linguists promotes sustainability of the documentation (cf. Arka this volume). Typically living in or near the communities, local linguists can closely interact with and supervise community members in their documentation efforts. If linguists are locals, they also usually possess greater sociocultural knowledge, expertise, and skills than foreign linguists. Such knowledge and skills are critical for the success and the sustainability of documentation activities. Second, financially, employing locals would be more cost effective than employing foreigners, and the money saved could be allocated for local needs.

However, very few local Indonesians have the necessary expertise and capacity of their international counterparts. Local linguists typically cannot compete to win international grants for language documentation. Among eighteen language documentation projects in Indonesia funded by the ELDP and the *Dokumentation Bedrohter Sprachen* (DOBES or in English, Documentation of Endangered Languages)⁶ over the past fifteen years, most of them have been proposed and led by foreign linguists. As the only Indonesians who have won ELDP grants to date (2018), we (both authors of this paper) attribute our achievements to our high-quality education in Australia and international collaborative research experience.

We believe that the locals' low capacity to compete internationally is a complex problem due to a combination of factors, such as the poor quality of education at all levels (including primary and secondary education and tertiary-level training in linguistics in particular) and insufficient financial support. In fact, the problem arguably started even earlier, socioculturally due to the oral tradition of (rural) societies where written literacy has been simply absent. Foreign linguists doing fieldwork in Indonesia are typically capable scholars who have been highly trained and have won competitive grants. They receive strong financial support and institutional backing from their home countries. Unsurprisingly, they are better equipped for fieldwork compared with local linguists. They are also in a better position to generate publications, organize seminars, and facilitate training activities. In contrast, local Indonesian linguists and linguistic students are typically not fortunate enough to receive proper training, financial backing,

⁴<https://archive.mpi.nl/>

⁵<https://lat1.lis.soas.ac.uk/ds/asv/?0>

⁶www.dobes.mpi.nl

and institutional support that could have equipped them to be as capable as foreign scholars. They are therefore disadvantaged by their lack of equal opportunity to win competitive documentation grants. Thus, for more local Indonesian linguists to win international grants and be able to manage and lead research teams, it is important that their expertise, capacity, and experience be improved through quality training and international collaborative research. These capacity and support issues are areas where foreign academics can be of immense help by contributing to capacity-building efforts at different levels, including tertiary education, as discussed in the next section.

4. Indonesian universities: Linguistic programs and advanced training

Reflecting on the need for capable local linguists brings us to the issue of advanced training in linguistics at both regional and national levels. Linguistic programs have been opened in major universities in Eastern Indonesia, including Hasanuddin University (Makassar, South Sulawesi), Cenderawasih University (Jayapura, Papua), and the University of Papua (Manokwari, Papua Barat), with the program at Hasanuddin University being the oldest, founded in the 1980s. Many of the faculty members in these linguistic programs have received doctoral training in linguistics overseas. Most of them subsequently become university administrators, overloaded with non-linguistic responsibilities and having almost no time to do proper fieldwork.

Additionally, while many linguistic graduates have been produced by the universities in Eastern Indonesia, the expertise of local graduates does not appear to be at par with the international standard. This complex issue is in fact part of a general education problem at all levels, including primary and secondary education, particularly in Eastern Indonesia. More generally, among 72 countries, Indonesia ranked 62nd in the Program for International Student Assessment results in 2015.⁷ If we take the set of national university rankings in Indonesia as an indicator, the universities in Eastern Indonesia are trailing behind their counterparts in Western Indonesia, particularly in Java. In Webometrics,⁸ the rankings of Indonesian universities show Universitas Pattimura Ambon in Maluku in the 73rd place nationally, while Universitas Nusa Cendana Kupang in Nusa Tenggara Timur (NTT) and the University of Papua occupy the 98th and the 99th positions, respectively. Indonesian universities are in turn lagging behind their international counterparts in the Asia-Pacific region and the world. According to the Quacquarelli Symonds university ranking, Universitas Indonesia (ranked first in Indonesia) is in the 277th place globally; Institut Teknologi Bandung (ranked second in Indonesia) is ranked 331st in the world, and Universitas Gajah Mada (ranked third in Indonesia) belongs to the 401–410 range in the world.⁹ Based on these relative rankings and our personal assessments about Eastern Indonesia, linguistic training in Eastern Indonesian universities needs improvement in various areas, including basic descriptive linguistics, typological-theoretical linguistics, fieldwork expertise, and modern language documentation. Contextualized training, specifically in preparing students for fieldwork in Eastern Indonesia, is also necessary and further discussed in the next section.

Foreign linguists working in Indonesia are typically aware of its low standard of tertiary training in linguistics. For this reason, they have collaborated with local universities to help improve the quality of tertiary education. For over two decades, from the 1980s to the 2000s, SIL linguists were based at Hasanuddin University and

⁷<https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf>

⁸www.webometrics.info/en/Asia/indonesia%20

⁹<https://www.topuniversities.com/university-rankings/world-university-rankings/2018>

involved in teaching linguistics under its graduate linguistic program. As mentioned, Nikolaus Himmelmann and his team helped set up the CELD at the University of Papua, with funding from the DOBES. The CELD has developed a training program focused on language documentation, integrating it with core courses in linguistics (phonology, morphology, and syntax) as part of the curriculum since 2009. From 2006 to 2018, short intensive training programs on language documentation and/or linguistic fieldwork have been organized by other foreign and local linguists. These include the DOBES-sponsored training in Bali, which was held twice (2006 and 2007) and led by Nikolaus Himmelmann; an ELDP-funded language documentation workshop conducted by Mandana Seyfeddinipur at Udayana University in Bali; and a series of workshops in Kupang, NTT, led by Asako Shiohara and Yanti (2017 and 2018) and funded by the Tokyo University of Foreign Studies grants. The success of such training programs is not easy to measure, however. The workshops seemed to have inspired some of the participants, including the first author of this paper, who then applied for grants/scholarships to conduct linguistic research and language documentation.

Three related points are noted here. First, all of the training programs mentioned have been sponsored by foreign funds. Second, they have been initiated and led by foreign scholars (with local collaboration). Third, the Indonesian educational system has an ongoing problem with leadership (cf. Arka this volume) and funding. A pressing issue is how to encourage more active participation in local linguistic programs, universities, and governments, individually and personally as scholars and collectively as institutions. The Indonesian government is currently putting pressure on academics and universities to conduct research and publish articles in respected international journals. The government has promised to provide more support for research and publication. President Jokowi's current administration has focused on building physical infrastructure for economic reasons but has promised to shift to the development of human resources in 2019, which is encouraging. We hope that this initiative will gradually make a difference in the quality of tertiary education in Eastern Indonesia, which will ultimately benefit linguistic fieldwork and language documentation in the region.

5. Working with local communities and contextualized training Successful linguistic documentation is determined by a combination of the fieldworker's capability and the local community's participation. The two components are interrelated. We reflect on the former in relation to contextual training, which should pave the way for the latter; for a discussion on participation issues in Indonesia, see Arka (this volume).

As discussed earlier, the capability issue is related to the inadequacy in high-level tertiary training. We believe that *contextual* training should be an essential component. This means that we must equip trainees (students, lecturers, language activists, etc.) with specific knowledge and practical skills to collect and analyze data for purposes that are relevant to fieldwork and language documentation in a given area. In the context of trainees from Eastern Indonesia, this means that the training should be contextualized to develop methodology and analytical skills targeting the familiarity with salient linguistic features and issues of the Austronesian and the Papuan languages of this region. They should also be trained in ethnographic skills and the knowledge of local cultures in Eastern Indonesia. In terms of the necessary tools, contextual training also entails practical hands-on courses in using modern devices that are specifically required for

language documentation (e.g., sessions on using annotation software, such as ELAN,¹⁰ Toolbox, and FLEX¹¹) and even the development of simple skills, such as how to save and back up data regularly on a computer. Our assessment is that the current curricula of the linguistic programs offered at universities in Indonesia do not generally include contextual knowledge and skills. Training programs for linguistic fieldwork and language documentation of the type organized in Manokwari and Kupang (as mentioned earlier) are therefore highly needed because they focus on developing such skills.

Working with local communities in Eastern Indonesia is more complex than language documentation alone. We often have to deal with different expectations of linguists and language communities so that both parties will be pleased and language documentation can proceed. We also have to handle local rivalries among people or clans. The expectations have been frequently associated with mutual benefits, with the local community or clan members often anticipating financial gain in return for their data given to non-local researchers or collaboration with the latter. This issue has arisen, largely due to past corrupt practices in government projects in which the Indonesian term *proyek* (project) has been equated with government money given to locals without accountability. While linguists fully understand that the local community members must be appropriately compensated for their time and efforts, we are also concerned about accountability in relation to project goals and outcomes, such as the amount of data to be collected, literacy materials to be produced, and outreach activities.

6. Final remarks In this final section, we reflect on the issue of different expectations between linguists and local communities. These issues are important for the local community members' active collaboration in documenting their own languages. As mentioned earlier, the problem with expectations has emerged due to past corrupt practices (associated with the term *proyek* in Eastern Indonesia) by government officials, non-government organizations, and other developmental agencies. As mentioned earlier, *proyek* has been misconstrued by locals as receiving "happy and easy money" (i.e., given by the government through various projects to compensate people without accountability). Local collaborators often tend to think that everyone involved in the project must receive regular payment throughout the duration of the project. Unfortunately, international developmental agencies, such as World Wide Fund for Nature with their conservation projects and UNICEF with their developmental projects, have used the same approach in their activities. Its negative effect is that local communities generalize their assumption that all outsiders come for *proyek* with a lot of cash to be distributed. They then often measure collaboration in terms of how much cash is given to local community members. The success of the project, particularly in relation to intangible outcomes, such as increased awareness of language endangerment, new skills in documentation, and literacy materials (e.g., for local elementary schools), is typically not their concern. It is a challenge to raise awareness that language documentation and maintenance are shared responsibilities and that while funds are needed in such efforts, it does not simply mean that a person can receive cash without an adequate contribution to the project, as the term *proyek* implies. For this reason, we suggest that foreign researchers avoid using the term *proyek* in describing their documentation research to locals. For instance, the CELD never uses the word *proyek* in its documentation activities. The terms

¹⁰<https://tla.mpi.nl/tools/tla-tools/elan/>

¹¹<https://software.sil.org/fieldworks/>

program and *aktivitas* 'activity' are more suitable because they carry a positive meaning, avoiding unwanted expectations from locals who seem to question project accountability.

Documentation projects come with a clear set of goals, planned activities, and a carefully considered budget. Thus, the limitations often constrain the kinds of activities and the number of local people recruited for particular documentation activities. In our experience, unexpected inquiries from locals include requests for payment in the form of food supplies, school supplies, clothes, and local infrastructure, all of which are not typically budgeted in the original proposal. Unfamiliarity with the documentation plan, budget limitation, and project accountability by the local community has often led to misunderstandings and different expectations; if not handled properly, these can derail documentation projects.

To avoid such misunderstandings, apart from avoiding the term *proyek* (as explained), one strategy that we have found useful is to adopt a persuasive participatory approach. Specifically, we first approach and invite local leaders (e.g., clan chiefs, village heads, educated locals) to be involved, seeking their advice to maximize community participation throughout the project and avoid certain possible problems. A community meeting involving all local stakeholders in the beginning of a documentation project is also essential. Such a meeting provides non-local researchers with the opportunity to openly explain the project goals, activities, and expectations. Based on our experience, the participatory approach mitigates conflicts of interest, allowing collaborative work between non-local researchers and local language community members, thus bringing the project to a fruitful conclusion. For this reason, a session on the participatory approach is part of the language documentation training offered in the CELD.

Our experience in the CELD has confirmed that involving local community members in language documentation leads to successful outcomes. In its first decade, CELD has supported local community members who have played active roles in documenting their languages (e.g., Woi, Iha, Mpur, and Yali). Certain local members have been trained and involved in working at the CELD. They have also carried out fieldwork in their own communities. Their deep understanding of their local cultures (a familiarity not shared by outsiders) has proven to be essential in the success of the CELD projects in Papua. From early on and throughout the documentation process, the CELD maintains close communication with the local communities to provide them with a proper understanding of the project goals and activities (e.g., documentation of their languages and cultures is part of promoting these). Open recruitment appears to be helpful as well (e.g., local members of the documentation project are appointed by the community through a fair and open process of consultation among community members). In return, the project's shared benefits for the whole community must be fairly negotiated.

To conclude, language documentation in Eastern Indonesia poses a challenge, whose success depends on several factors, including sound planning to anticipate a range of linguistic and non-linguistic problems. The local community's active participation is essential. In our experience, such engagement is not always easy because each speech community has its own local culture, needs, and expectations. These may vary considerably from one language and one place to another in Eastern Indonesia, and they are dynamic in nature. Thus, efforts to enable local communities' involvement in language documentation are ongoing and locally specific, in which all stakeholders (linguists, funding bodies, educational institutions, government institutions, and local communities) have to collaborate toward a common goal for the benefit of the local communities.


References

- Arka, I Wayan, J. Kosmas, & Nyoman I. Suparsa. 2007. *Bahasa rongga: Tatabahasa acuan ringkas*. Jakarta: Penerbit Universitas Atma Jaya (PUAJ).
- Arka, I Wayan. 2008. Local Autonomy, local capacity building and support for minority languages: Field experiences from Indonesia. *Language Documentation & Conservation Special Publication No. 1*. 66–92.
- Arka, I Wayan. 2015. *Bahasa rongga: Deskripsi, tipologi dan teori*. Jakarta: PKBB, Universitas Katolik Atma Jaya.
- Arka, I Wayan. 2016. Language documentation in Indonesia: Framing innovative linguistic research in the diversity of ethno-ecology context. A paper presented at *KIMLI 2016*, 24–27 August 2016. Universitas Udayana, Denpasar, Indonesia.
- Baird, Louise. 2002. *A grammar of Keo: an Austronesian language of East Nusantara*. PhD dissertation, The Australian National University.
- Bellwood, Peter. 1997. *Prehistory of the Indo-Malaysian archipelago*. Honolulu: University of Hawai'i Press
- Blust, Robert. 1993. Central and Central-Eastern Malayo-Polynesian. *Oceanic Linguistics* 32(2). 241–293.
- Blust, Robert. 2009. The position of the languages of Eastern Indonesia: A reply to Donohue and Grimes. *Oceanic Linguistics* 48(1). 36–77.
- Bowden, John. 2001. *Taba: Description of a South Halmahera language*. Canberra: Pacific Linguistics.
- Donohue, Mark & Charles E. Grimes. 2008. Yet more on the position of the languages of Eastern Indonesia and East Timor. *Oceanic Linguistics* 47(1). 114–158.
- Florey, Margaret. & Nikolaus P. Himmelmann. 2007. New directions in field linguistics: Training strategies for language documentation in Indonesia. In Margaret Florey (ed.), *Endangered languages of Austronesia*, 212–140. Oxford: Oxford University Press.
- Foley, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology* 29. 357–404.
- Gippert, Jost, Nikolaus P. Himmelmann, & Ulrike Mosel (eds.). 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.
- Hajek, John. 2010. Towards a phonological overview of the vowel and consonant systems of East Nusantara. East Nusantara: Typological and areal analyses. In Michael C Ewing & Marian Klamer (eds.), *East Nusantara: Typological and areal analyses*, 25–46. Canberra: Pacific Linguistics.
- Harmon, David. 1996. Losing species, losing languages: Connections between linguistic and biological diversity. *Southwest Journal of Linguistics* 15. 89–108.
- Harmon, David & Luisa Maffi. 2002. Are linguistic and biological diversity linked? *Conservation Biology in Practice* (3). 26–2
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Himmelmann, Nikolaus P. 2005. The Austronesian languages of Asia and Madagascar: Typological characteristics. In K. Alexander Adelaar & Nikolaus P. Himmelmann (eds.), *The Austronesian languages of Asia and Madagascar*, 110–181. London: Routledge.
- Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.

- Hisa, La, Agustinus Mahuze, & I Wayan Arka. 2017. The ethnobotanical-linguistic documentation of Sago: a preliminary report from Merauke. *Linguistik Indonesia* 35(2). 187–200.
- Holton, Gary & Marian Klamer. 2017. The Papuan Languages of East Nusantara and the Bird's Head. In Bill Palmer (ed.), *The languages and linguistics of New Guinea: A comprehensive guide*, 549–619. Berlin: Mouton de Gruyter.
- Klamer, Marian. 2002a. Ten years of synchronic Austronesian linguistics (1991–2002). *Lingua* (112). 933–965.
- Klamer, Marian. 2002b. Typical features of Austronesian languages in Central/Eastern Indonesia. *Oceanic Linguistics* (41). 363–384
- Klamer, Marian. 2008a. The languages of East Nusantara: state-of-the-art. Paper presented at the *international workshop on minority languages in the Malay/Indonesian speaking worlds*. Leiden, 28 June 2018.
- Klamer, Marian & Michael Ewing. 2010. “The languages of East Nusantara: An introduction”. In Michael Ewing & Marian Klamer (eds.), *East Nusantara: Typological and areal analyses*, 1–24. Canberra: Pacific Linguistics.
- Klamer, Marian & Frantisek Kratochvil. 2014. The expression of number in languages of East Nusantara: An overview. In Marian Klamer & Frantisek Kratochvil (eds.), *Number and quantity in East Nusantara*, 1–14. Canberra: Asia-Pacific Linguistics.
- Ndiken, Novita Y. S. 2011. *Verbal agreement in Malind*. Undergraduate thesis, Manokwari Universitas Negeri Papua.
- Olsson, Bruno. 2017. *The Coastal Marind Language*. Singapore: Nanyang Technological University dissertation.
- Platenkamp, J.D.M. 1990. North Hamahera: Non-Austronesian languages Austronesian cultures? Paper presented at the *Oosters Genootschap in Netherland*. Leiden, 23 May 1989.
- Ross, Malcolm. 1988. *Proto-Oceanic and the Austronesian languages of western Melanesia*. Canberra. Pacific Linguistics.
- Ross, Malcolm. 1995. Some current issues in Austronesian linguistics. In Darrell T. Tryon (ed.), *Comparative Austronesian Dictionary 1*, 45–120. Berlin: Mouton de Gruyter.
- Ross, Malcolm. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. In Andrew Pawley, Robert Attenborough, Jack Golson, & Robin Hide (eds.), *Papuan past: Cultural, linguistic and biological histories of Papuan-speaking peoples*, 15–65. Canberra: Pacific Linguistics.
- Sawaki, Yusuf W., Jeanete Lekeneny, Anna Rumakeuw & Sonja Riesberg. 2016. Language documentation and capacity building in West Papua: The Center for Endangered Languages Documentation, Universitas Papua. Paper presented at *Konferensi Internasional Masyarakat Linguistik Indonesia 2016*, 23–27 August 2016. Denpasar. Bali, Indonesia.
- Sawaki, Yusuf W. 2011. Empowering students and speech community members in language documentation activities. Paper presented at the *Symposium on Strengthening Language Maintenance Through Cooperative Training Strategies*. The University of Melbourne, Melbourne, Australia. 18–19 August 2011.
- Sawaki, Yusuf W. 2016. *A grammar of Woor, an Austronesian language of Yapen Island, Western New Guinea*. PhD dissertation, The Australian National University.

- Sawaki, Yusuf. W. 2016. Language documentation and conservation in Papua – what is it for? Paper presented at the *1st International Symposium on Biodiversity and marine Conservation of the Bird's Head Seascape*, 2–5 November 2016. Manokwari, Universitas Papua.
- Schapper, Antoinette. 2015. Wallacea, a Linguistic Area. *Archipel* 90. 99–151.
- Shiohara, Asako & Yanti. 2018. Training for documenting minority languages of Indonesia: Practice, benefit and challenges. Paper presented at *KOLITA 16*, 10–13 April 2018. Atma Jaya University, Jakarta, Indonesia
- Spriggs, Matthew. 1997. *The islands of Melanesians*. Oxford. Blackwell Publishers.
- Timmer, Jaap. 2002. A bibliographic essay on the southwestern Kepala Burung (Bird's Head, Doberai) of Papua. In *Papuaweb's Annotated Bibliographies*. Manokwari, Jayapura, and Canberra: Universitas Papua, Universitas cenderawasih dan Australian National University. (<http://www.papuaweb.org/bib/abib/jt-kepala.pdf>)
- Turvey, Samuel T. & Nathalie Pettorelli. 2014. Spatial congruence in language and species Richness but not threat in the world's top linguistic hotspot. *Proceedings of the Royal Society B - Biological Sciences* 281: 20141644. (<http://dx.doi.org/10.1098/rspb.2014.1644>)
- Wurm, Stephen A. 1982. *Papuan languages of Oceania*. Tübingen: Gunter Narr Verlag.

Yusuf Sawaki
yusuf.sawaki@anu.edu.au

I Wayan Arka
wayan.arka@anu.edu.au
 orcid.org/0000-0002-2819-6186

Reflections on linguistic fieldwork in Australia

Ruth Singer
Australia National University

Shifts in White-Indigenous relations started to re-shape relations between field linguists and Australian Indigenous communities from the 1970s. So well before Himmelmann (1998) appeared, linguists working on Australian Indigenous languages had been discussing topics such as ethical engagement with Indigenous communities, accessibility of recordings and the best use of technology in archiving and recording. After Himmelmann (1998) appeared, these topics emerged as key topics in language documentation which led to more of these kinds of discussions not only among Australian linguists but also with linguists around the world. The development of language documentation as a field of research fostered greater collaboration between Indigenous communities, linguists, researchers from other disciplines and technology specialists in Australia. New funding initiatives followed the publication of Himmelmann (1998), providing additional support for documentation projects on Australian Indigenous languages. Since the 2000's government support for Indigenous-led initiatives around language has declined in Australia. But growing support for Indigenous researchers within universities is enabling Indigenous communities to become more equal partners in research on their languages.

The emergence of 'language documentation' as a distinct subfield of linguistics undoubtedly had an influence on fieldwork in Australia¹. However, it is not easy to trace this influence among the other changes already in train when Himmelmann (1998) appeared. Fieldwork practices are influenced not only by developments within academia but also social change in Indigenous communities and the national context of White-

¹Acknowledgements: This paper draws on understandings gained through work done as part of a research partnership with Waruwi Community. Recently financial support has come from the Australian Research Council (DE140100232 and FL130100111), the Hans Rausing Endangered Languages Programme, the Centre of Excellence in the Dynamics of Language and the Research Unit for Indigenous Language. Thanks very much to Nick Thieberger, Isabel O'Keeffe and Jenny Green for discussing some of the topics in this article and to Rosey Billington for commenting on a draft. All opinions remain my own.

Indigenous relations. In Australia, the politics of 'Indigenous affairs' changed significantly in the 1970's, creating new possibilities for how (White) linguists and Indigenous communities might work together². The origins of contemporary fieldwork practice can be traced back to the intense, enthusiastic and creative collaborations between linguists and Indigenous people that took place at this time, the birth of the self-determination era. As Indigenous communities gained a stronger political voice, many directed community energies and funding towards supporting their languages. In parallel, descriptive linguists working with Indigenous communities began to reconsider their approach.

Wilkins' (1992) paper, 'Linguistic research under Aboriginal control', reflects on his experiences while working at an Indigenous-run bilingual school in the 1980's. The government funded a number of Indigenous bilingual school programs from 1973 onwards and these were the site of many productive engagements between Indigenous communities and linguists, many of whom started their work in the community as school teachers (Devlin et al. 2017; Laughren 2000). Wilkins identified significant tensions between the goals he had as a linguist, working on a grammatical description of the Mparntwe Arrernte language and the goals of the Indigenous community affiliated with that language. These are now focal topics in the field of language documentation: ethical approaches to working with communities and how best to create accessible documentation materials. Wilkins recounts how the Indigenous representative body, the *Aboriginal Languages Association* presented their statement on the 'Linguistic Rights of Aboriginal and Islander Communities' to the 1984 *Australian Linguistics Society (ALS)* meeting which then was accepted as ALS policy³. The activities that Wilkins describes illustrate the climate of postcolonial reflection of that time. Many linguists, like other White Australians at the time, had a strong desire for reconciliation with Indigenous Australians.

Since the 1970's, some Indigenous communities have been able to employ linguists like Wilkins, drawing on government funds and mining royalties. Linguists employed by Indigenous-run organizations are answerable to the Indigenous community first, rather than to academia and thus have more motivation than others to reconsider their fieldwork practices. Many field linguists have been employed outside of academia as expert witnesses in land claims, as interpreters, in Indigenous schools, arts centres and in language centres. Language Centres are a key meeting point for academic linguists and Indigenous communities. Although they were more numerous in the past, language centres continue to provide employment for linguistics graduates, many of whom return to universities at some point, bringing with them a more collaborative approach to working with Indigenous communities (Sharp & Thieberger 2001). At a recent University of Melbourne symposium that brought together researchers and their Indigenous research partners, Indigenous scholar Sana Nakata made a comment that people in Indigenous communities spend a lot of valuable time training White people in how to work with their communities. While linguists have always tried pass on understandings gained in the field to their students, the emergence of 'language documentation' has seen these topics recognized as a part of academic research proper and discussed widely in academic literature. This literature may help new field linguists to understand the basis of good collaboration with Indigenous communities and lessen the burden on Indigenous

²In this paper I use the term 'linguist' to refer to the mainly White linguists who do fieldwork on Indigenous languages and are employed by universities rather than Indigenous communities. Other linguists are also crucial to fieldwork on Australian languages.

³See the statement at: <https://als.asn.au/AboutALS/Policies> (accessed 10/10/18).

communities that Nakata mentioned. However as Thieberger has pointed out “The tension between the academic research agenda and the desires of speakers nevertheless remains and requires constant reflection and negotiation” (Thieberger 2016 p.91).

Long before the political developments of the self-determination era there was a strong tradition of linguistic fieldwork in Australia. Involving the creation of grammars, dictionaries and text collections, it was very much in step with the Boasian tradition in the United States. A decline in fieldwork on less documented languages is said to have occurred there with the rise of Chomsky’s research program in the 1960’s (Woodbury 2010). In fact this was the time when linguist Ken Hale visited and fired up a generation of young Australian linguists to go out and do basic descriptive work (Simpson et al. 2001). This influx of new linguists seems to coincide with an increase in the quality of analyses found in descriptive grammars, as well as their depth and breadth (see for example Tsunoda 1974; McKay 1975; Dixon 1977). During this time, linguists also paid attention to comparing phenomena between Australian languages which helped to identify widespread Australian phenomena such as ergativity that had often been underanalysed.

From the 1970’s onwards there is a clear increase not only in the quality of descriptive materials that were being produced but also the quantity. The Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS) employed linguist Jeffrey Heath to do descriptive work on Indigenous languages from 1973-1977. His publications, produced with speakers of a number of south-east Arnhem Land languages, are highly valued by their descendants, most of whom have not been able to learn these languages as children. His most detailed work is on Wubuy (Nunggubuyu); comprising a grammar, dictionary and text collection interconnected by such comprehensive cross-referencing that they have been described as a pre-digital hypertext (Musgrave & Thieberger 2012). There is a clear sense of the rapid loss of languages among linguists who did fieldwork on Australian languages in the 1970’s because so many worked with the last fluent speakers of a language or language variety. The idea of ‘language endangerment’ struck a chord with these linguists, drawing parallels with the sharp decline in Australia’s biodiversity since 1788, among the fastest rate of extinction worldwide.

The developments of the 1970’s, created a receptive audience for Himmelmann’s (1998) ‘language documentation’ manifesto among linguists doing fieldwork on Indigenous Australian languages. At the time it appeared, there were many field linguists working in Indigenous Australian communities with the aim of creating lasting records of Indigenous languages. The appearance of Himmelmann (1998) was accompanied by plenty of debate among these linguists as to whether ‘language documentation’ was a highly innovative idea or simply a new way of looking at existing practices (Woodbury 2010). With hindsight however, these debates seem less relevant as ‘language documentation’ has taken on a life of its own. It has become a banner under which field linguists have organized themselves and worked on ways to better meet the needs of Indigenous communities. The field of language documentation has played an important role in fostering new collaborations between linguists, other disciplinary specialists, Indigenous communities and technology experts.

Himmelmann played a key role in developing the Dokumentation Bedrohter Sprachen (DoBeS)⁴ program funded by the Volkswagen Foundation which funded two language

⁴see: <http://dobes.mpi.nl/>

documentation projects in Australia⁵. The Iwaidjan languages project was focussed on less-documented languages spoken on the Cobourg Peninsula in western Arnhem Land, Northern Territory. The DoBeS funding supported an interdisciplinary team that included a number of linguists as well as an anthropologist and a musicologist. This meant that close academic collaborations begun in the field, formed the basis for later analysis, publication and archiving. Funding rules were flexible enough to support a field linguist stationed in one Indigenous community for a number of years. This kind of placement greatly aids collaboration with communities and gives much greater scope for supporting communities to develop their own capacity to do linguistic research.

The emergence of philanthropic initiatives such as DoBeS and the Hans Rausing Endangered Languages Programme (HRELP) helped pave the way for Australian universities to recognize language documentation projects as research projects. Short-term project funding became available through the Australian Research Council (ARC) in the early 2000's. Early on, a number of language documentation projects were funded, both team projects and individual research fellowships. The ARC only funds projects that are 'innovative' so language documentation as a new idea helped attract more funding for fieldwork. HRELP funded dozens of projects on Indigenous Australian languages. Together, these diverse sources of funding have made it possible for linguists to respond directly to community-identified goals and document language together with sign, gesture, narrative practice, drawing, music, plant and animal knowledge and complex kinship systems. While linguists tend to focus on language alone, Indigenous communities often want to preserve and maintain holistic Indigenous knowledge systems requiring an interdisciplinary approach involving musicologists, historians, archaeologists, biologists and anthropologists.

One goal of the new language documentation paradigm was to make documentary materials available in an accessible manner and it quickly became apparent that digital language archives were the best way to do this. Making materials accessible is key to making linguistic fieldworkers more accountable to the communities they research, academia and the general public (Berez-Kroeker et al. 2018). Documentary materials in Australian Indigenous languages have long found a safe home in the archive of the Australian Institute for Aboriginal and Torres Strait Islander Studies. The Institute has enshrined in its constitution, an obligation to serve the interests of Australia's Indigenous people. However, it was slow to respond to the promise of digital archiving. Looking for a way to create a digital archive of their materials, linguists turned to the PARADISEC⁶ digital archive, and it has become an important place for digital records of Australian languages. The archive is located across three Australian universities (University of Melbourne, Australian National University and Sydney University) and the outreach activities of founders Nick Thieberger and Linda Barwick have ensured that few researchers of Australian Indigenous languages, music or dance remain unaware of the benefits of digital archiving. PARADISEC has also provided training in recording techniques, compiling metadata, annotation in Elan, etc. to fieldworkers all around the country. An important part of the work of field linguists in the past few decades has been returning early archival materials back to communities. After enrichment of these

⁵A number of research projects on Australian languages were based at the Max Planck Research Institutes in Nijmegen and Leipzig at this time, which further strengthened German-Australian collaborations around language documentation.

⁶<http://www.paradisec.org.au/>

materials through further fieldwork and annotation they are then re-archived, improving the digital record of the language (Thomas & Neale 2011; Harris 2014).

Australian linguists embraced digital technologies during the nineties and naughties just when key aspects of language documentation were emerging. Many took up Toolbox (SIL International 2018), then Elan (2018) and more recently they have been involved in creating apps to aid documentation in the field such as Aikuma (Bird et al. 2014) and Kinsight (Foley 2017). The latter app (still in development) comes out of a concerted investment into new technology for language documentation by the Transcription Acceleration Project, Centre of Excellence in the Dynamics of Language (CoEDL). However, new developments in technology invariably raise new ethical concerns. For example in Arnhem Land, the widespread ideology of language ownership, whereby languages are owned through one's father, coupled with the distribution of authority (i.e. who can speak for each language) across complex kinship networks, can make checking permissions for materials no simple task.

I will look at two issues in particular, raised by new technology. The first is the way that participants and their lives are much more visible in recent recordings. Traditionally, linguists recorded word lists, isolated sentences and often also a few texts. Within the language documentation paradigm there is an emphasis on 'natural' language use; language in its natural context and also a concern with covering a diverse range of kinds of speech events (Seifart 2008). This has led to linguists recording conversations, multi-participant narratives, songs, sand stories and different kinds of verbal art. As it became easier to record for many hours at a time, simultaneously on multiple recording devices, linguists began to record entire speech events, such as informal conversations or performances of music and dance. If the context is a natural context for language use, then participants are doing something else as well as making a recording, and are less likely to be thinking about the fact that somebody could later watch the recording. In longer recordings, the sheer quantity of data can make it harder for participants to audit all recordings in which they appear. The details of people's lives become much more visible in these kinds of richer and more rewarding recordings. In many ways this is a good thing, as disembodied language data is less valuable to everyone. However it does raise issues about how exposed participants become and how much control they have over this.

The second issue that technology has raised, that will be discussed is the closely related issue of accessibility. The digital archive makes it possible to access recordings much more easily, which can be very useful for speakers and linguists. The greater accessibility of recordings means, however, that more detailed discussions are needed with participants about how to handle the recordings. Where recordings are made available online, linguists often find themselves in the difficult position of discussing access restrictions with elders who have little sense of how truly accessible something can be once it is online. Indigenous Australian communities have always been quite concerned about the circulation of recordings of their people's image and voice. Traditionally, consideration for the bereaved meant avoiding exposing them to images and recordings of the deceased for some time after the death. Images and recordings of the living were also treated with care, in case they got into the wrong hands. However, since phones brought their digital cameras and audiovisual recording capabilities into every home, protocols in Indigenous communities seem to be changing. Images are now widely distributed through social media, including images of the deceased. For example, G. Yunupingu, an internationally known musician from Eastern Arnhem Land died last year before a film about his life was

released. His family decided to approve the release of the film without any delay, which was said to be because they felt it was important to get his message out. This is just one family's decision; strict protocols still hold for the most part. However there is a need to continually re-evaluate community views about the accessibility of archived materials, as participants that appear in recording sessions pass away and also as views about how to handle images and recordings change in Indigenous communities.

The HRELP became a key source of funds for language documentation projects on Australian languages after the DoBeS program ended and the ARC stopped funding projects focussed purely on language documentation. However, a few years ago, HRELP introduced an strong open access policy which meant some linguists doing fieldwork with Indigenous Australian communities no longer consider HRELP a suitable source of funds. The requirements are that most recordings made in current HRELP-funded projects are set to the 'O' (open) setting in the ELAR online archive, although a small proportion may be kept closed if this is well justified by the depositor. For many Indigenous communities it would be difficult to make this kind of commitment in the grant application stage. That said, a number of Indigenous Australian projects have still been funded since the adoption of HRELP's open access policy. One of these is a language documentation project on the Kunbarlang language, which is spoken mainly by older people in western Arnhem Land.

Dr Isabel O'Keeffe is lead investigator of the Kunbarlang language project⁷ and Professor Linda Barwick and I are co-investigators. O'Keeffe and I held meetings with the remaining Kunbarlang speakers and their descendants before applying for HRELP funds. They were not concerned about the open access conditions of HRELP funding and expressed a real urgency to make recordings available as widely as possible so that their young people could hear them. At this stage the project is going well and people are happy with the level of access to recordings. However it is hard to know whether things will change in the future if Kunbarlang is no longer spoken. The fact that we have proceeded at each step with openness and many Kunbarlang-affiliated people have relatively high levels of digital literacy, gives us reason to hope that the grant conditions will not cause problems down the track. In effect, the project is an experiment in digital archiving and accessibility as well as a language documentation project. However, it would put less pressure on linguists doing fieldwork on Australian languages if funding bodies gently encouraged open access rather than making it a condition of funding.

Himmelman (1998) ushered in an era of many new funding initiatives for endangered languages, only a few have been mentioned here. However, the heyday of innovative international funding programs for language documentation has clearly passed. Regardless, language documentation is surviving peak popularity to become an established part of linguistics. Language documentation can now be a central component (if not the sole element) of a linguist's career. Successful language documentation projects and their outputs can be listed on a CV. Recognising language documentation as valid research makes it easier for linguists to devote time to it. Efforts have been made to get the Australian university 'publications accounting' system to count corpora and dictionaries as research outputs just like journal articles (Thieberger et al. 2016). Some Australian universities now recognize dictionaries of Indigenous languages as research monographs, formerly they were not counted as such. Language documentation is also a part of many ARC-funded projects although there is usually a specific research focus to the project such as child language development, small-scale multilingualism or the processing of

⁷Full title: Empowering Indigenous youth to create a comprehensive pan-varietal, ethnobiological, anthropological record of Kun-barlang through training in low-cost language documentation technology.

polysynthetic languages. The language documentation that is part of these projects is broadly valuable, as these projects investigate phenomena in languages that are not well documented.

This paper began by looking at how Indigenous Australians got their languages on the national and international agenda in the 1970's, in a climate of global concern about language endangerment. This inspiring period, known as the self-determination era, ended around 2007⁸ by which point government policy had clearly shifted from supporting self-determination towards more assimilationist policies that constructed Indigenous culture including language as a barrier to Indigenous development. Since 2007 we have seen the closure of most Indigenous bilingual schools, reduced support for Indigenous language and culture programs in all schools and a reduction in funding for Indigenous language projects⁹. However, there is a growing understanding of the value of Indigenous knowledge systems within universities and there are more Indigenous students and staff at universities than ever before. The ARC's Discovery Indigenous grant program has funded a five year fellowship for Indigenous scholar Elizabeth Marrkilyi Ellis to document contemporary verbal arts in the western desert region, together with a team of linguists and anthropologists. Universities are developing pathways for greater inclusion of Indigenous research partners in the research process, through funding for Indigenous research partners to visit universities and travel to conferences and co-present papers. New generations of linguists have a better developed sense of responsibility to the Indigenous communities they work with. So we can anticipate more of an emphasis on linguists being accountable to the communities they work with in the future. Language documentation has provided a framework for field linguists in Australia to discuss important issues and push for change.

⁸A suite of changes are generally seen by scholars of Indigenous politics as marking the end of the self-determination era. The start of these changes are dated by the 2007 Northern Territory Emergency Response (NTER) 'The Intervention' and include the abolition of the national Indigenous representative body ATSIC, the closure of most government-run bilingual schools and the replacement of Indigenous community councils with super-shires (Altman 2016).

⁹Although in 2018 this has been increased.


References

- Altman, Jon. 2016. Imagining Mumeka: Bureaucratic and Kuninjku perspectives. In Nicolas Peterson & Fred Myers (eds.), *Experiments in self-determination: Histories of the outstation movement in Australia*, 279–300. Canberra: Australian National University Press.
- 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. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.
- Bird, Steven, Florian R. Hanke, Oliver Adams & Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–5. (<http://acl2014.org/acl2014/W14-22/index.html>) (Accessed 2018-10-10)
- Devlin, Brian, Samantha Disbray & Nancy Devlin (eds.). 2017. *History of bilingual education in the Northern Territory: People, Programs and Policies*. Singapore: Springer.
- Dixon, Robert MW. 1977. *A grammar of Yidiny*. Cambridge: Cambridge University Press.
- ELAN. 2018. [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. (<https://tla.mpi.nl/tools/tla-tools/elan/>)
- Foley, Ben. 2017. *Kinsight* [Computer software]. Brisbane: Centre of Excellence in the Dynamics of Language. (<https://play.google.com/store/apps/details?id=com.cbmm.kinsight&hl=en>)
- Harris, Amanda (ed.). 2014. *Circulating cultures*. Canberra: Australian National University Press.
- Himmelman, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.
- Laughren, Mary. 2000. Australian Aboriginal languages: Their contemporary status and functions. *Handbook of Australian languages*, Volume 5, 1–19. Oxford: Oxford University Press.
- McKay, Graham Richard. 1975. Rembarnga: A language of central Arnhem Land. Canberra: Australian National University dissertation.
- Musgrave, Simon & Nick Thieberger. 2012. Language description and hypertext: Nunggubuyu as a case study. In Sebastian Nordoff (ed.), *Electronic grammaticography* (Language Documentation & Conservation Special Publication 4), 63–77. <http://hdl.handle.net/10125/4530>
- Seifart, Frank. 2008. On the representativeness of language documentations. In Peter K. Austin (ed.), *Language Documentation and Description*, vol. 5, 60–76. London: SOAS. <http://www.elpublishing.org/>.
- SIL International (2018) *Toolbox* [Computer software] <http://www.sil.org/computing/toolbox/>
- Sharp, Janet & Nicholas Thieberger. 2001. Wangka Maya, the Pilbara Aboriginal Language Centre. In Jane Simpson, David Nash, Mary Laughren, Peter K. Austin & Barry Alpher (eds.), *Forty years on: Ken Hale and Australian languages*, 325–335. Canberra: Pacific Linguistics. doi: <https://doi.org/10.2307/3623459>.
- Simpson, Jane, David Nash, Mary Laughren, Peter K. Austin & Barry Alpher (eds.). 2001. *Forty Years On: Ken Hale and Australian Languages*. Canberra: Pacific Linguistics. doi:<https://doi.org/10.2307/3623459>

- Thieberger, Nick. 2016. Documentary linguistics: Methodological challenges and innovatory responses. *Applied Linguistics* 37(1). 88–99.
- Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21.
- Thomas, Martin & Margo Neale (eds.). 2011. *Exploring the legacy of the 1948 Arnhem Land Expedition*. Canberra: Australia National University Press.
- Tsunoda, Tasaku. 1974. A grammar of the Warungu language, North Queensland. Melbourne: Monash University thesis.
- Wilkins, David. 1992. Linguistic research under aboriginal control: A personal account of fieldwork in central Australia. *Australian Journal of Linguistics* 12(1). 171–200. doi:10.1080/07268609208599475.
- Woodbury, Anthony C. 2010. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.

Ruth Singer

ruth.singer@anu.edu.au

 orcid.org/0000-0003-4915-3262

In search of island treasures: Language documentation in the Pacific

Alexandre François
LaTTiCe, CNRS–ENS–Sorbonne nouvelle
Australian National University

The Pacific region is home to about 1,500 languages, with a strong concentration of linguistic diversity in Melanesia. The turn towards documentary linguistics, initiated in the 1980s and theorized by N. Himmelmann, has encouraged linguists to prepare, archive and distribute large corpora of audio and video recordings in a broad array of Pacific languages, many of which are endangered. The strength of language documentation is to entail the mutual exchange of skills and knowledge between linguists and speaker communities. Their members can access archived resources, or create their own. Importantly, they can also appropriate the outcome of these documentary efforts to promote literacy within their school systems, and to consolidate or revitalize their heritage languages against the increasing pressure of dominant tongues. While providing an overview of the general progress made in the documentation of Pacific languages in the last twenty years, this paper also reports on my own experience with documenting and promoting languages in Island Melanesia since 1997.

1. Approaching language documentation in the Pacific This paper reflects on twenty years of linguistic documentation in the Pacific. After an overview of the region's rich linguistic ecology (§1), I will survey the progress made so far in documenting and archiving valuable recordings from the region (§2). Crucial to the success of language documentation is also its relevance to speaker communities in their strive to preserve and revive their own languages (§3).

1.1 Overview of Pacific languages The Pacific region is home to about 20% of the world's languages (Simons & Fennig 2018), and hosts a great number of different language

families. In terms of human and linguistic geography, the term *Pacific* commonly refers to the set of inhabited islands located within the Pacific Ocean, south of the 30° N parallel. Depending on the context, the term may also include the Philippines and Indonesia—usually considered part of SE Asia—as well as Australia; but these areas are covered by other chapters in this volume (see Arka & Sawaki (2018) for SE Asia; Singer (2018) for Australia in this volume). The present chapter will thus identify the Pacific (Figure 1) as the vast area defined by the three subregions of Melanesia, Micronesia and Polynesia.¹

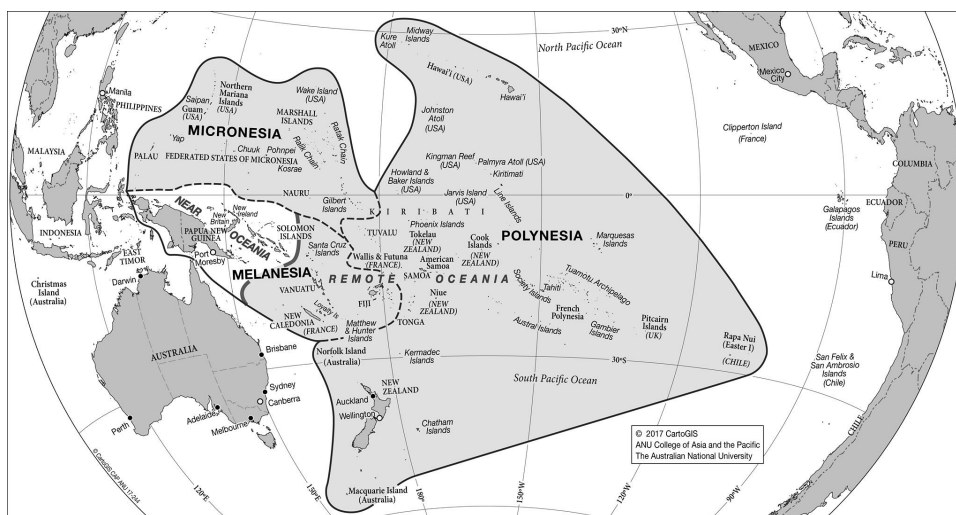


Figure 1: A map of the Pacific region, showing the subdivision into Melanesia, Micronesia, Polynesia; and the archaeological divide between Near Oceania (white) and Remote Oceania (gray) [© ANU College of Asia and the Pacific, CartoGIS, 2017]

The names for these three subregions were first introduced by Dumont d’Urville in 1831, and do not reflect accurately the populations’ prehistory (Green 1991, Tcherkézoff 2009). More recently, Pawley & Green (1973) proposed to divide the same region based on archaeological criteria, into two areas:

- *Near Oceania* (consisting of the island of New Guinea and most of the Solomon Islands) was first settled by *Homo sapiens* more than 50,000 years ago;
- *Remote Oceania* (the rest of the Pacific—see grayed area in Figure 1) was only settled during the last 4,000 years.

As Map 1 shows, the boundary between Near and Remote Oceanic splits apart the area traditionally labelled as “Melanesia” (Green 1991).

The two areas also differ in their linguistic make-up. The more recently settled Remote Oceania features only one family, namely Austronesian—more exactly, the *Oceanic* branch of the Austronesian phylum. As for Near Oceania, it is home to about 80 genealogically unrelated language families and isolates,² making it the world’s genealogically most

¹Map provided by CartoGIS Services, ANU College of Asia and the Pacific, The Australian National University.

²“The Papuasphere [...] contains, by the current count, 862 languages comprising 43 distinct families and 37 isolates.” (Palmer 2018:6).

diverse area. These various families are collectively referred to using the umbrella term *Papuan*, to which one must add a more recent layer of Austronesian migrations.

Table 1 shows the unequal distribution of indigenous languages across the different subregions of the Pacific.³ Micronesia and Polynesia, which were settled relatively recently, show lower linguistic diversity, with only about 60 languages together for such a vast area. By contrast, the region of Melanesia shows considerable diversity with a sheer total of 1419 languages—whether Austronesian (557) or non-Austronesian (862).

Area	Language family	# languages
MELANESIA, Near Oceania (New Guinea, Solomon Islands)	“Papuan” (≈80 families)	862
MELANESIA, Near Oceania (New Guinea, Solomon Islands)	Austronesian	345
MELANESIA, Remote Oceania (eastern Solomon Island, Vanuatu, New Caledonia, Fiji)	Austronesian (Oceanic)	212
POLYNESIA	=	38
MICRONESIA	=	21
Pacific region		1478

Table 1: Distribution of languages across the different subregions of the Pacific

1.2 Language density and vitality According to Simons & Fennig (2018),⁴ Pacific languages, considered as a whole, show an average of 5,271 speakers per language, with a median value of 980. These figures are the lowest of all continents: they can be compared, respectively, with the world’s average of 1 million speakers per language, with a median value of 7000.

One can in fact observe extreme discrepancies in language density between different areas of the Pacific (Pawley 1981, 2007). On one extreme, the language with most speakers is Sāmoan, with 413,000 speakers (*Ethnologue*). On the other extreme, the average size of a language community in Vanuatu at the beginning of the 20th century—at a time when the archipelago went through a demographic bottleneck—was “as low as 565 speakers per language” (François et al. 2015: 9). This goes to show the drastic gap in the language ecology across different parts of the Pacific—as “Melanesian diversity” (Dutton 1995, Unseth & Landweer 2012) contrasts so strongly with “Polynesian homogeneity” (Pawley 1981).

The diversity of Papuan languages may arguably be explained by the considerable time depth of human settlement in Near Oceania, and a long history of migrations and language evolution. But what is perhaps more striking is that even the Austronesian-speaking populations, which have had less than four millennia of *in situ* development, achieved

³Numbers are taken or inferred from the database Glottolog 3.2 (Hammarström et al. 2018). Usual disclaimers apply when counting languages. Also, note that “38” is the number of languages belonging to the Polynesian branch of Oceanic languages: only about half of these are spoken within the *Polynesian triangle* (the area labelled ‘Polynesia’ on Figure 1), while the remainder, known as *Polynesian outliers*, lie geographically in Micronesia or Melanesia.

⁴See <https://www.ethnologue.com/statistics>.

a similar rate of language density. Vanuatu, for example, was first settled 3,100 years ago by speakers of Proto Oceanic (Bedford & Spriggs 2008); and in that relatively short time-span, the archipelago's small population (currently 0.3 million) managed to diversify into 138 distinct languages—making it the country with the world's highest linguistic density *per capita* (François et al. 2015).

1.3 Different landscapes, different strategies This overview of the varying linguistic landscapes found across the Pacific (§1.1) entails quite different approaches when it comes to language documentation.

The languages with larger speaker populations numbering over 100,000, such as those found in major Polynesian centers (Sāmoan, Tongan, Māori, Tahitian, etc.) are certainly threatened in the long term due to the pressure of colonial languages—French, English—and of globalizing trends; but for the immediate future, they can be deemed safe from immediate endangerment. Because these languages have already been the object of grammatical or lexical descriptions, the work of linguists is rather to document the various styles and registers of these languages—whether that be technical vocabulary, verbal art, poetry (e.g. Meyer 2013 for Tahitian)—or the internal dynamics of their variation (Love 1979, Duranti 1997 for Sāmoan).

The situation is different with the many languages of Melanesia, or indeed with the demographically smaller languages of Micronesia or Polynesia. About half of Pacific languages are spoken by populations below the threshold of 1,000 speakers, which makes them more vulnerable to the risk of language shift and loss. In view of the sheer number and diversity of smaller languages of the Pacific, the urgent task is often for linguists to describe and document the linguistic practices of these speech communities while the languages are still vital. The last two decades have seen considerable effort in that direction; and while a lot remains to be done in the region, it is already possible to report on various successful endeavors in the domain of language documentation in the Pacific. The next sections will illustrate some of these efforts, and outline ways in which they can be appropriated by language communities.

2. Fieldwork archives

2.1 From description to documentation A few decades ago, the work of missionaries and pioneer linguists typically consisted in collecting basic data in the form of wordlists (e.g. Leenhardt 1946 for New Caledonia, Tryon 1976 for Vanuatu) or grammar sketches (e.g. Codrington 1885, Ray 1926 for Melanesian languages). Apart from translations of the Scriptures, it was rare to collect or publish texts, or other samples of connected speech.

The tide turned when linguists understood that their role was to record languages in the way they were actually spoken. Rather than eliciting wordlists or translating grammarians' sentences, language describers began conscientiously collecting high-quality data. This involves recording spontaneous speech in various forms: narratives, procedural texts and explanations, personal memories, conversations. That important evolution had already begun in the 1970s—as witnessed, for example, by the volumes of stories collected in various languages of Melanesia (e.g. Ozanne-Rivierre 1975–79; Paton 1979; Bensa & Rivierre 1982; Facey 1988) and Polynesia (e.g. Frimigacci et al. 1995). In the same spirit, researchers and engineers at Paris-based CNRS–LaCiTO created the first online

audio archive in endangered languages, as early as 1996 (Jacobson, Michailovsky & Lowe 2001), bringing together valuable fieldwork recordings with their text annotations.⁵

Himmelman's explicit proposal for *language documentation* (Himmelman 1998) thus came at a timely point in the evolution of researchers' practices. Rather than giving value to the sole results of linguistic analysis in the form of academic papers, grammars or dictionaries, the focus was shifting towards the quality and availability of actual samples of spontaneous speech in the various languages under study.

The rationale for the new focus on high-quality linguistic data was manifold. The insistence on gathering spontaneous speech serves an aesthetic criterion—the wish to pay tribute to the world's intangible linguistic heritage—but also a scientific one. If linguistic description and typology are meant to be an empirical science, it is not sufficient to translate pre-hashed questionnaire sentences derived from a linguist's theoretical model; instead, it is essential that the object of our scrutiny exists in the form of an observable corpus, independent of any aprioristic prejudice on what sort of structures we should expect to find.

2.2 Language archives Pacific languages are well represented in a number of online archives. The OLAC *Language Resource Catalog*⁶ lists several collections featuring Pacific resources. Some of these archives, like *Rosetta* or *SIL-LCA*, include simple wordlists, grammar sketches, or translated sections of the Scriptures. Only some of the archives here mentioned pertain to documentary linguistics strictly speaking, in the sense of linguistic corpora based on naturalistic speech. The form usually taken by these language resources is as audio or video recordings, of varying length, typically ranging between 1 and 20 minutes each. These media are archived either as raw sound recordings, or as sound enriched with text annotations: transcription and free translation—with the additional possibility of interlinear glossing.

Table 2 lists the number of audio recordings⁷ featured in the catalog of the world's four main language archives dealing with the Pacific: PARADISEC (Thieberger & Barwick 2012); DOBES and other databases hosted by the MPI (Brugman et al. 2002, Wittenburg et al. 2002); CNRS–LaCiTO's *Pangloss Collection*⁸ (Jacobson et al. 2001; Michailovsky et al. 2014); and University of Hawai'i's *Kaipuleohone* (Albarillo & Thieberger 2009; Berez 2013). I only list resources in indigenous languages of the Pacific, including pidgins and creoles, to the exclusion of colonial languages.

Recordings from Pacific languages are more or less prominent in each archive. For example, the 3039 Pacific audio samples found in the MPI Language Archive correspond to a mere 5 percent of their entire catalog of recordings. By contrast, Pacific languages

⁵LaCiTO's online archive, which was later named the *Pangloss Collection* (Michailovsky et al. 2014), was initially developed by Boyd Michailovsky, Martine Mazaudon and John Lowe, and later expanded by Michel Jacobson. Pangloss is the largest collection within the CoCoON repository—see fn.8.

⁶OLAC Language Resource Catalog: <http://dla.library.upenn.edu/dla/olac/>. Note that ELAR, the *Endangered Languages Archive* developed at SOAS (Nathan 2010), is not featured under OLAC, and is thus unfortunately absent from the present statistics; yet that archive contains 36 archival deposits from the Pacific. Particularly noteworthy is Mike Franjeh's collection on Northern Ambrym languages (Franjeh 2018), which earned DELAMAN's Franz Boas 2019 award for best online multimedia documentary collection.

⁷This corresponds to the type 'Sound' among the categories proposed by the *Dublin Core Metadata Initiative* used by OLAC. Similar statistics could be carried out with video (DCMI 'MovingImage'), yet in this paper I will restrict myself to audio resources, for the sake of brevity and consistency.

⁸In the OLAC catalog, the *Pangloss Collection* appears under the name "CoCoON" (*Collections de Corpus Oraux Numériques*), a compilation of several audio archives hosted by CNRS' Huma-num infrastructure. LaCiTO's Pangloss is the largest collection within CoCoON, and the only one dealing with Pacific languages.

Archive	Institution	# Audio in catalog	# Audio from Pacific	% Catalog
PARADISEC	U Melbourne, ANU, U Sydney; CoEDL	9748	4458	46 %
DOBES, MPI-PL, Language collections	MPI for Psycholinguistics (Nijmegen)	66721	3039	5 %
<i>Pangloss Collection</i>	CNRS–LaCiTO	3302	1422	43 %
<i>Kaipuleohone</i>	U Hawai‘i	2571	804	31 %

Table 2: Language archives displaying audio resources in indigenous Pacific languages

are proportionally better represented in PARADISEC (which has *Pacific* in its very name) and in the *Pangloss Collection*.

Archives differ in their precise geographical coverage. As Table 3 shows, languages of Papua New Guinea (whether Papuan or Austronesian) are well represented in PARADISEC and MPI-DOBES. Those of Vanuatu are mostly found in PARADISEC and Pangloss. Pangloss is also the place to go for languages of New Caledonia. Kaipuleohone’s strong spots are PNG and Micronesia.

	PARADISEC	MPI- DOBES	Pangloss	Kaipu- leohone	Total
Melanesia: Pidgins & creoles	147	25	13	2	187
PNG, Solomons: <i>Papuan</i>	1953	1806	–	126	3885
PNG: <i>Austronesian</i>	861	686	–	354	1901
Solomons: <i>Austronesian</i>	459	178	93	33	763
Vanuatu	874	113	869	–	1856
New Caledonia	27	–	404	3	434
Fiji	88	–	–	50	138
Micronesian	17	–	–	212	229
Polynesian	32	231	43	24	330
TOTAL	4458	3039	1422	804	9723

Table 3: Number of media resources for each archive, organized by geographic and linguistic area

2.3 A sample of individual corpora Table 4 lists the ten richest audio corpora—judging by the number of media resources—for individual languages of the Pacific. The four main archives (cf. Table 2) are represented, as well as the four main countries making up Melanesia.

Among these online corpora, those hosted by the MPI, Kaipuleohone or Paradisec, are difficult to access as they require special authorization, even for consultation. The most readily accessible resources are those of the *Pangloss Collection*, which makes it easier to

Language	Family	Country	Linguist	Archive	# audio
<i>Yeli Dnye</i>	(Papuan)	PNG	S. Levinson	MPI	722
<i>Mwotlap</i>	Oceanic	Vanuatu	A. François	Pangloss	504
<i>Savosavo</i>	(Papuan)	Solomons	C. Wegener	MPI	462
<i>Saliba</i>	Oceanic	PNG	A. Margetts	MPI	423
<i>Bebeli</i>	Oceanic	PNG	H. Sato	Kaipuleohone	356
<i>Titan</i>	Oceanic	PNG	T. Schwartz	PARADISEC	327
<i>South West Bay</i>	Oceanic	Vanuatu	L. Dimock	PARADISEC	302
<i>Blablanga</i>	Oceanic	Solomons	R. Voica	PARADISEC	292
<i>Teop</i>	Oceanic	PNG	U. Mosel	MPI	234
<i>Cèmuhî</i>	Oceanic	New Caledonia	JC Rivierre	Pangloss	231

Table 4: The ten richest audio corpora for individual languages from the Pacific (source: OLAC)

find out about statistics. I will give a brief overview of the Cèmuhî and Mwotlap corpora available there.

Cèmuhî, a tonal language of New Caledonia, is represented on Pangloss by 231 audio resources, recorded in the field between 1965 and 1979 by the late Jean-Claude Rivierre.⁹ Each entry is a wav file, digitized from legacy reel-to-reel tapes. Most recordings feature traditional narratives, whether myths or folktales (see Rivierre & Ozanne-Rivierre 1980, Bensa & Rivierre 1982). Among these 231 audio entries, 56 are accompanied by text annotations, consisting of a transcription, a free translation, and glosses—see Figure 2.

Mwotlap, a language of the Banks Islands (north Vanuatu), is featured in 504 audio recordings, which I collected between 1997 and 2011. With about 52 hours in total, the Mwotlap corpus forms about half of the recordings I have archived on Pangloss. Altogether, these consist of 962 resources, totalling 104 hours of sound,¹⁰ in twenty-three different languages: four Oceanic languages of the Solomon Islands (Lovono, Tanema, Teanu, Tikopia), eighteen Oceanic languages of Vanuatu (Araki, Dorig, Hiw, Koro, Lakon, Lehali, Lemerig, Lo-Toga, Löyöp, Mota, Mwerlap, Mwesen, Mwotlap, Nume, Olat, Vera'a, Volow, Vurës), and one creole (Bislama).

My archives on Pangloss take the form of sound files with metadata, downloadable under a Creative Commons licence (see fn.11). About 105 resources (including 38 for Mwotlap) are accompanied by time-aligned transcriptions and other text annotations, similar to Figure 2 above.

About one third of the archived resources are musical performances of various sorts, from dances to sung poetry: these formed the basis of a discographic publication together with the ethnomusicologist Monika Stern (François & Stern 2013). The remaining two thirds (69%) represent connected speech—mostly folk narratives (389 stories), but also procedural explanations, conversation, elicitation. A selection of narratives was the

⁹Besides Cèmuhî, Rivierre has also archived recordings in three other Kanak languages: Paicî, Numèè and Bwato—see <http://tiny.cc/Rivierre-archives>.

¹⁰Link: <http://tiny.cc/Francois-archives>.

The screenshot shows the 'Pangloss Collection' website interface. At the top, there are logos for CNRS, LaCiTO, CoCoOn, Huma-Num, EFL, and ANR. Below the logos is a navigation bar with 'The Pangloss Collection', 'Corpus access', 'Dictionaries', 'Submit resources', and 'Help'. A search bar is on the right. The main content area displays three entries:

S16 [audio icon] **kā mē ē tēbwō jē-da ēni hwà ò ē-jè kàapo hě té-kō cilè pā-li nàhì- dāame kā mē ē nie opé kā ē āli- ē-jè iké ne pwō- li wīnāado**
 et quand elle est assise ci-en haut ici dans maison elle qui celle-ci Kaapo alors que en train de veiller sur le-déf. petit de-
 dāame kā mē ē nie opé kā ē āli- ē-jè iké ne pwō- li wīnāado
 chef et quand elle regarde en bas par ici et elle voit- celle-ci iké au dessus de- les déf. vivres
 Kaapo est dans la case en train de s'occuper du petit et, de chez elle, aperçoit iké sur le tas de vivres.

S17 [audio icon] **kā ē tēmèhì-èng**
 kā ē tēmèhì-èng
 et elle reconnaît-elle
 Elle reconnaît sa soeur,

S18 [audio icon] **kā ē cē tēbwō mwo ò ē-jè kā ē tēē é**
 kā ē cē tēbwō mwo ò ē-jè kā ē tēē é
 et elle .. est assise ..dès lors elle qui celle-ci et elle reste à pleurer
 fond en larmes et n'arrête pas de sangloter.

Figure 2: Screenshot of the Cémuhi corpus: The traditional story “Les écailles de poisson de Tiwécaalè” is presented in a time-aligned transcription, with translation and glossing (Jean-Claude Rivierre, LaCiTO–CNRS; story by Bernadette Tyèn).

source of several booklets I created towards the consolidation of vernacular literacy in various speaker communities (§3.5).

3. Community outreach Language documentation not only involves the work of linguists, but also favors initiatives by speaker communities towards the preservation and revitalization of their heritage languages.

3.1 Enhancing access to the resources The audio or video resources made by linguists can be highly valued by the community of speakers. These documents preserve the memory of specific individuals, storytellers or singers or personalities who can now be remembered by their relatives, descendants and countrymen. The recordings also encapsulate cultural knowledge, folk traditions, oral history, important narratives and artistic forms that deserve to be passed on to the next generations. Finally, they also capture the various shapes taken by a living language, whether in the form of dialogues, stories, or verbal art; obviously, this linguistic testimony is all the more precious when the language is threatened with extinction within a few decades or years. For all these reasons, the current speakers of the language, or their descendants to come, constitute a key audience for our documentary work (Turin et al. 2013).

One aspect to be developed in the near future is the ease with which community members can access our archives, even when they are not technically savvy, or familiar with linguists' circles. Archives on a particular language should turn up in public search engine results, and at least the catalog of resources be easily and intuitively searchable. Most archives require the end user to create an account, and often to ask for permission to access specific resources: while this may be necessary in some cases, this is often a *de facto* obstacle to many community members who would like to casually listen to their heritage language without having to go through the intimidating process of an official request. In this sense, a fully public option certainly has its advantages.

LaCiTO's Pangloss Collection (or its sister CoCoON: see fn.8) is the only online archive that seems to fill those requirements at the moment: even if its interface could still be made more appealing to a lay audience, at least it is intuitive enough that anyone can easily navigate it, retrieve some resources, and listen to them right away—since they are all provided in Open access¹¹ and require no authorization.

One important advantage of this easy access is the possibility to share specific recordings with community members, e.g. via social media. In the last years 2014-2017, I created Facebook pages for the communities speaking respectively the languages of Araki, Mwotlap, Hiw (Vanuatu) and Teanu (Solomon Islands). Besides promoting the use of vernacular languages in writing (§3.5), each page also gives us the opportunity to share links to individual archived fieldwork recordings. Occasionally, I can send the link to a resource as a reply to a member's specific request—whether they're looking for a particular story, or an old song, or a sample of their grandfather's voice. The possibility to access recordings so readily was always warmly welcomed by group members. This is a simple and efficient way to return the fruit of our research to the younger members of the communities, in a way meaningful to them.

3.2 Searching across corpora and archives Documentary resources on Pacific languages are currently distributed across different archives, each with its own interface, principles, technical options. A certain level of inter-operability across archives has already been achieved through the common adoption of the standards of *Dublin Core Metadata Initiative*, and the shared connection with the OLAC initiative (Simons & Bird 2003). This has made it possible to cross-search several archives at once based on geographical or linguistic criteria, and navigate from one multimedia repository to another, across institutional boundaries.

However, at the moment this cross-integration among archives remains limited, and would benefit from being enhanced. Suppose an anthropologist, or a Pacific islander, would like to scan all existing archives from a certain region of the Pacific, say for stories containing the terms “canoe”, or “shark”, or “magic” (either in the title, or in the text's English translation). At the moment, many individual repositories lack a search function or a concordance tool, which is definitely a gap to fill. A community member, or a scholar, may wish to search one or several text corpora for a certain word form or gloss—a simple endeavor that is often still impossible. Such a search engine could also make it possible to look for an important placename, or to retrieve a valuable story which had ceased to be transmitted. The title proposed in the metadata—often the only searchable segment—is not always sufficient. Ideally, the tools for navigating or interrogating corpora would be pooled together (technology permitting) across different archives.

¹¹ The default license at Pangloss is CC BY-NC-ND 3.0 (Attribution, Non-commercial, No Derivatives). Recordings that are sensitive for cultural or social reasons are not displayed publicly on Pangloss.

3.3 Repatriating recordings Like in most places of the world (Pearce & Rice 2013), the internet in the Pacific is nowadays less and less accessed through computers, and more frequently through mobile interfaces; this may well inform our practices in terms of designing our tools in the future. Yet in spite of fast improvements, in many rural areas of the Pacific internet access still remains costly and unreliable, so much that offline solutions are still welcome for the diffusion of knowledge.

In 2011, I thus chose to repatriate all my field recordings to speaker communities in the form of a local digital copy—first, to the Vanuatu Cultural Centre in Port Vila; and secondly, to a newly created *Torres-Banks cultural centre* on the island of Motalava (François 2012; Michailovsky et al. 2014: 131). Vanuatu’s Alliance Française helped maintain this cultural centre, funded a laptop, and provided technical training to local curators.

Admittedly, searching through a thousand audio recordings in 23 languages, from many locations, with so many genres and contributors, would constitute a challenge for the local users, most of whom had never used a computer. For that reason, I designed an intuitive way to search through the archives. I exploited the possibilities offered by free media players such as iTunes or Winamp, then available on local interfaces. Each archived recording was converted from wav into an mp3 file. The latter was then enriched with id3 metadata, which were automatically imported from Dublin Core descriptors as already stored in the Pangloss archive: e.g. the name of the speaker or storyteller became the “ARTIST”, the place of recording was mapped onto the “GROUPING” tag, and so on. Each recording was associated with a photograph—generally, a portrait of the speaker—which took the place of an album’s artwork. The result was an enticing multimedia digital library featuring hundreds of recordings, that could be explored either through pictures, or through a multifaceted search involving: *language; title; name of speaker or musician; date of recording; location; genre; duration* (François 2012).¹² This local search interface was welcomed by the community members, who were able to search through the collection in many ways, and retrieve recordings of their interest on the computer—whether samples of speech or music. They would then download these recordings from the public media library to their mobile phones used as offline mp3 devices, and distribute them to their friends and families. This proved a successful way to share digital resources amongst the community, even in an offline context.

Back in 2011, I found it technically difficult to replicate the same rich search interface using online tools. However, given the current spread of the internet, and the latest developments of mobile applications, it should now become easier to come up with user-friendly search tools so as to distribute the fruit of our documentary work to non-academic users through mobile-based devices.

3.4 Community-driven documentation We just saw how communities can benefit from the efforts of language documentation carried out by linguists, through increased access to valuable recordings. Interestingly, these results can also inspire the speakers themselves to pursue the work of documentation on their own language.

While the description of a language’s grammar or lexicon requires solid academic training in the domain of linguistics, the documentation of linguistic practices is a different sort of endeavour, whose crucial ingredients include: familiarity with the language to be documented, and ability to transcribe it; acquaintance with the cultural universe attached

¹²The video at <https://youtu.be/hZGm0CLzxU8> demonstrates the potential of that interface.

to it; understanding of what is at stake in language endangerment and documentation; personal motivation. When these conditions are present, native speakers are often in a good position to bring about language documentation themselves. They can then set out to record the language under its various manifestations—whether traditional narratives, procedural texts, conversations—using audio or video technologies (Carew et al. 2015, Bettinson & Bird 2017).

Recent technological developments have made access easier to decent-quality microphones, including through the use of commercial smartphones. Some apps have been developed for the creation of dictionaries—such as *Ma! Iwaidja* for the Iwaidja language of northern Australia, or *Ma! BenaBena*, with respect to a Papuan language of Papua New Guinea (Birch 2013). Their interface allows crowdsourcing, and the enrichment of lexical data by the community members themselves (Carew et al. 2015: 314).

Another tool tailored for community-driven documentation is the Android app “Aikuma” (Bird et al. 2013; Bettinson & Bird 2017), itself superseded by *LIG-Aikuma* (Blachon et al. 2016). This user-friendly application allows native speakers to enrich an existing recording, or a newly-created one, using vocal annotation, such as slow-speech “respeaking” or translation. The interface is designed to be used intuitively by community members even if they are not literate. The software was first used with speakers of Usarufa in Papua New Guinea, a language where *aikuma* means “meeting place”.¹³

More recently, *Aikuma* has also become the name of a collaborative project, with the aim to promote the celebration of indigenous languages through oral performances of storytelling and verbal art. The project has been active online: <https://twitter.com/AikumaProject>.

3.5 Language learning and revitalization Among the many positive outcomes of language documentation projects, is their possible usefulness for language revitalization and language learning.

Through collaborative workshops and initiatives, web-based projects and team efforts, an increasing number of activities are taking place across the Pacific—like elsewhere in the world—that promote the use of vernacular languages in speaking and in writing. One could cite the cases of Māori *kōhanga reo* or “language nests” in New Zealand (Benton 1989, King 2001); of *pūnana leo* for Hawaiian (Warner 2001); or similar attempts in French territories, whether French Polynesia (Paia 2014) or New Caledonia (Moyse-Faurie 2012, Vernaudon 2015)... These revitalization projects are only tangentially related to documentary linguistics per se (see Penfield & Tucker 2011), but they participate in a general push to increase the exposure of younger generations to their legacy languages in the variety of their manifestations. In some cases, efforts in language learning and revitalization are directly linked to the enterprise of linguistic documentation, and with the recording of spontaneous speech from fluent speakers.

Indeed, the high-quality samples of fluent speech, such as the audio and video resources found in archives, deserve to be exploited for their teaching potential. The target audience may be outsiders wishing to learn a new language; or members of the language community—whether speakers or semi-speakers themselves—wishing to access valuable recordings in their heritage languages. Complete beginners would first need to have access to teaching materials; but once they’ve acquired enough fluency to understand

¹³See <http://www.aikuma.org/faq.html>

simple stories, then the audio or video recordings—especially if enriched with annotations and glosses—are of considerable help to increase their linguistic competence.

Accessing the audio or video recordings has a great learning potential, especially for learners who wish to maintain fluency in their heritage language, by hearing high-quality storytelling from their elders. But another way in which language documentation can be used in a teaching context, is by developing skills in literacy, whether in regular schools or in community-led learning groups.

While many possible examples could be cited, I will briefly report on my personal experience in Melanesia. The majority of the 23 languages on which I carried out language documentation in Vanuatu and the Solomons (§2.3) lacked any stable orthography when I began working on them. One of my roles, as a linguist, has been to sort out the phonology of each language (François 2011a: 194), and design a set of spelling conventions, which I then discuss with community members during public meetings. Once a system has been agreed upon, I can start sharing the transcriptions of my field recordings with speakers. In many cases, this exchange proves a pivotal moment for the community, whose oral language is finally endowed with an “official” orthography.

The next stage is to create teaching materials to make sure community members can master the spelling system that was agreed upon. The effort required for this learning depends a lot on the difficulty of the language’s phonology—particularly, its vowel inventory. Because Teanu, the main language of Vanikoro (François 2009), has only five vowels /i e a o u/—the same as those in the Roman script—its speakers master the orthography very fast. But it takes more time, and more exposure to written materials, in order to transcribe consistently such languages as Hiw with its nine vowels /i ɪ e ə a ɯ o ɔ/; or Lemerig with eleven /i ɪ ε æ a œ ø ɸ ɔ ʊ u/.

Over the years, I have produced a number of booklets for literacy education. All volumes are also available on my homepage in digital format (François 2015)—a total of 21 books so far, of two different kinds. Ten books took the form of a basic alphabet primer, exemplifying each grapheme with a selection of words and phrases, with rich homemade illustrations. The remaining eleven volumes are story books, and constitute readers for a more advanced level of literacy. Their list is provided in Table 5.

Apart from one reader whose text I wrote in Mwotlap with the help of Edgar Howard (François & Howard 2000), all volumes showcase a selection of traditional narratives in each language, taken from my corpus of transcribed stories. Their length ranges from 36 to 78 pages each. They are monolingual, as they aim to encourage literacy in the vernacular language, as opposed to the country’s dominant languages of education (French, English or Bislama) which are often people’s default choices when it comes to writing. Figures 3 and 4 illustrate one page from two of those story books.

The literacy materials in question were initially self-published and self-funded by my family. In 2006, the Vanuatu bureau of Alliance Française added their support to six of these volumes, by funding the printing of 300 copies, as well as their shipping to different local schools in the Banks group; we renewed our collaboration again in 2015, with respect to four books destined to the communities of the Torres islands. Since then, I have heard numerous reports that these literacy materials have been successfully perused in various schools of the Torres-Banks province of Vanuatu—whether this was an initiative of the local teacher, or an application of the country’s recent pledges to develop vernacular literacy programs in early school years (Vanuatu Ministry of Education 2012).

Language	Area	Year	Book title	Contents
Araki	Espiritu Santo	2008	<i>Sorosoro māran Raki</i>	Stories in Araki (south Santo)
Dorig	Gaua, Banks	2011	<i>O susrig ble mraw vata Dōrig</i>	Stories in Dorig
Hiw	Torres Is.	2015	<i>Ne vegevage fōssē 'n Mefavtit</i>	The legend of the hero Megravtit
Lemerig	Vanua Lava, Banks	2006	<i>Nvāv 'ām 'a Lēmērig</i>	Stories in Lemerig
Lakon	Gaua, Banks	2011	<i>Suusuu pule maraw avan Lakon</i>	Stories in Lakon
Lo-Toga	Torres Is.	2015	<i>Ne vegevage sē te Lō mi ne Toge</i>	Stories from Lo and Toga islands
Mwesen	Vanua Lava, Banks	2004	<i>O olñevu ta turmō ta Mēsēn</i>	Stories in Mwesen
Mwotlap	Banks	2003	<i>Tog tog i van en</i>	'Once upon a time': Stories in Mwotlap
Mwotlap	Banks	2000	<i>Bulsal, dam galsi me lēklek</i>	'Follow me, my friend': Language reader
Olrat	Gaua, Banks	2011	<i>Ususraa pule maraw men Ōlrat</i>	Stories in Olrat
Teanu +	Temotu, Solomon Is	2012	<i>Liatevo iepiene nē piene akapa</i>	Stories in the three languages of Vanikoro: Teanu, Lovono, Tanema

Table 5: Text materials produced by the author based on his documentary work, aimed at literacy development and language revitalization

Now that the internet has reached speaker communities (since 2010 in the capital, since 2017 in some rural areas), I have witnessed increased use of vernacular languages in writing—whether in mailing lists, in texting, or on Facebook (§3.1). For some languages that used to show inconsistent attempts at transcription, it appears that younger contributors now show more consistency in their spelling. Some explained to me how they learned to write their own language using the literacy materials I produced, which gave them more confidence when writing their own legacy language.

52

Ba kê nemyös so kê mas wuh köyö, namyös nonon so kê so niwuh köyö.

Kê magal köyö van so köyö vanvan nen e tō köyö mitiy, wa tō kê nivan hōw, tō wuh matmat köyö mi naqyēn.

Tō— kê nihatig hag nen tō— mahag van; kê net van so “Ohoo! Yoge gōh et mitimiy te!” Kê menyē van, menyē van, tateh.

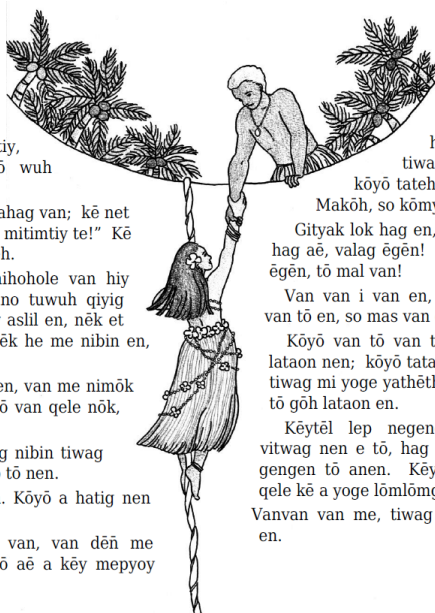
Mālmal nonon Wētamat en nihohole van hiy nōlōmgep e wo “Imam mino tuwuh qiyig dōyō. Ba nēk valag hiy yow aslil en, nēk et van nibin lol tō anen, ba nēk he me nibin en, ba lep me noel.”

Kê nilep nibin tiwag mi noel nen, van me nimōk veteg köyö van, vēshēt köyö van qele nōk, köyö hohole!

“Hatig tō, dō van!” Mōk veteg nibin tiwag mi noel tō köyö hohole noyō tō nen.

Mahē nognog meyen me ēgēn. Köyö a hatig nen tō, lep namtehal tō a— van.

Köyö van van van van va— van, van dēn me lemtehal a nagayga dam tō aē a key mepyoy tēqēl tō kê aē en.



Wētamat yōnteg a so “Ooh! Vētmahē kê ninyen ēagōh!”, so kê nivan wuh vege köyö a köyö nēh en. Kê nivan hōw me qele kê, nibin tiwag mi noel hohole tō nen, köyö tateh. “Aah! Kōmyō so gal no? Makōh, so kōmyō ave?!”

Gityak lok hag en, lep nohos nonon, yow kal hag aē, valag ēgēn! Yoge mal dam ketket alge ēgēn, tō mal van!

Van van i van en, nēdēmdēm noyō so köyö van tō en, so mas van dēn a lataon.

Köyö van tō van tō va— van, van dēn hōw lataon nen; köyö tatal kēkēl laptō, imam nonon tiwag mi yoge yathēthēn en, mal van me, tō hag tō gōh lataon en.

Kēytēl lep negengen nagaytēl den nisto vitwag nen e tō, hag tō aslil lefranda nan nen, gengen tō anen. Kēytēl gengen en, so et van qele kê a yoge lōmlōmgep en nivanvan van me! Vanvan van me, tiwag mi mālmāl non Wētamat en.

Figure 3: A page from a story book in Mwoṭlap, Motalava I., Vanuatu (François 2003)



Wōrō ēn hag va— van, naw tē 'āt avōh, wōrō tē rākās.

Rākās mongēē wāl-wāl neñ, Qat gēn tē mākā to sala nē ga to' jen nē.

Nē 'n hālāqāg tāgāh mērē heg, Tañro tē hālāqāg kēl talvōn um, nē tē hālāqāg talvōn.

Va—van, we tē wāhā kere vata nagē Tañro neñ mēn too qet ga tutun we nē tē ginteg.

Ginteg hōw neñ, tutuan pah nok makē raṃos.

Nē tē row vēgēn hag, nē gēn tē luwluw gēn ek, sa “Mok raṃos! Na ga marēs nēk a rēg! Nēk a vutgi!”

We raṃos neñ tē rēg, tē vutgi ajew.

Gēē tē vēgēn suu-

32

Suusuu pule maraw avan Lakon

suu. Tañro tē vēgēn suusuu gēē ma.

Gēē 'n vēgēn vēgēn va—n, Qat sa: “Mok raṃos! Nēk rēg, nēk vutgi lēh hag!”

Raṃos neñtē vutgi sa tē vutgi kakarānkē vata hag, hag neñ to lē pōlō.

Gēē 'n jēn hag neñ, sa “Ha'ēh! Mok raṃos! Nēk et kaōl!”

We 'n raṃos neñ tē kaōl.

Kaōl kaōl kaōl va—van, jēn hōw lē vanō neñ, nē tē hālāqāg jen, tutuan tē hālāqāg, nē tē gih lekteg uhli raṃos neñ.

Tañro neñ nē 'n kakal ma, va rigtāg we nē tē ukāg raṃos, raṃos tē 'āqā wutā i Qasval.



Storian long lanwis blong Lakon

33

Figure 4: A page from a story book in Lakon – Gaua I., Vanuatu (François 2011b)

4. Conclusion: A mutual benefit for linguists and communities The movement of documentary linguistics, as it emerged in the 1990s and was theorized by Himmelmann (1998), has meant a leap forward in the quality standards of the primary data serving as the basis for language description and analysis. In the Pacific region, this progress has endowed numerous endangered languages with rich corpora of naturalistic speech, in the form of audio and video recordings, with the frequent addition of text annotations. Future years should increase the mutual integration of language documentation and description, so as to reinforce the accountability of linguistic analyses based on solid empirical data.

Yet the relevance of language documentation goes beyond providing empirical material for the linguist. By focusing on naturalistic speech in different social contexts, the archives produced by the documentary enterprise have major potential for a broad array of audiences—anthropologists and historians, experts of oral literature and ethnomusicologists, language teachers and learners. Most importantly, the endeavor of documentary linguistics is proving of high relevance to members of Pacific speaker communities, whether their wish is to hear the voices of their elders, to create and peruse literacy materials in their legacy languages, or to hand over their linguistic and cultural knowledge to the upcoming generations.

References


- Albarillo, Emily E., & Nick Thieberger. 2009. Kaipuleohone, University of Hawai'i's Digital Ethnographic Archive. *Language Documentation & Conservation* 3. 1–14.
- Bedford, Stuart & Matthew Spriggs. 2008. Northern Vanuatu as a Pacific crossroads: The archaeology of discovery, interaction, and the emergence of the “ethnographic present.” *Asian Perspectives* 47. 95–120.
- Bensa, Alban, & Jean-Claude Rivierre. 1982. *Les Chemins de l'Alliance. L'organisation sociale et ses représentations en Nouvelle-Calédonie.* (Langues et Cultures du Pacifique.) Paris: Société d'Etudes Linguistiques et Anthropologiques de France.
- Benton, Nena. 1989. Education, language decline and language revitalisation: The case of Maori in New Zealand. *Language and Education* 3(2). 65–82.
- Berez, Andrea. 2013. The Digital Archiving of Endangered Language Oral Traditions: *Kaipuleohone* at the University of Hawai'i and *C'ek'aedi Hwnax* in Alaska. *Oral Tradition* 28(2). 261–270.
- Bettinson, Mat, & Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 156–164. Honolulu, Hawai'i, USA, March 6–7, 2017.
- Birch, Bruce. 2013. The Ma! Project: Crowdsourcing software for language documentation. In Amanda Harris, Nick Thieberger & Linda Barwick (eds.), *Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures* <http://hdl.handle.net/2123/9858>
- Bird, Steven, Florian Hanke & Haejoong Lee. 2013. Collaborative language documentation with networked smartphones. In Amanda Harris, Nick Thieberger & Linda Barwick (eds.), *Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures* <http://hdl.handle.net/2123/9857>
- Blachon, David, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker & Annie Riailand. 2016. Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, May 2016, Yogyakarta, Indonesia. *Procedia computer science*.
- Brugman, Hennie, Stephen Levinson, Romuald Skiba & Peter Wittenburg. 2002. The DOBES Archive: Its Purpose and Implementation. In Peter K. Austin, Helen Dry & Peter Wittenburg (eds.), *Proceedings of the international LREC workshop on resources and tools in field linguistics*. Paris: European Language Resources Association (ELRA). *Proceedings of the international LREC workshop on resources and tools in field linguistics*. Paris: European Language Resources Association (ELRA).
- Carew, Margaret, Jennifer Green, Inge Kral, Rachel Nordlinger & Ruth Singer. 2015. Getting in touch: Language and digital inclusion in Australian indigenous communities. *Language Documentation & Conservation* 9. 307–23.
- Codrington, Robert H. 1885. *The Melanesian languages*. Oxford: Clarendon Press.
- Duranti, Alessandro. 1997. Indexical speech across Sāmoan communities. *American Anthropologist* 99: 342–354.
- Dutton, Thomas Edward. 1995. Language contact and change in Melanesia. In Peter S. Bellwood, James J. Fox & Darrell Tryon (eds.), *The Austronesians: Historical and comparative perspectives*, 207–228. Canberra: Australian National University.

- Facey, Ellen E. 1988. *Nguna voices: text and culture from central Vanuatu*. Calgary: University of Calgary.
- François, Alexandre. 2003. *Tog tog i van en* ['Once upon a time...']. Collection of stories from the oral tradition of Mwotlap. Monolingual in Mwotlap, for local use. Illustrated by Sawako François. Self-published. 78 pp.
- François, Alexandre. 2009. The languages of Vanikoro: Three lexicons and one grammar. In Bethwyn Evans (ed.), *Discovering history through language: Papers in honour of Malcolm Ross* (Pacific Linguistics 605), 103-126. Canberra: Australian National University.
- François, Alexandre. 2011a. Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence. *Journal of Historical Linguistics* 1(2), 175-246.
- François, Alexandre. 2011b. *Suusuu pule maraw avan Lakon* ['Traditional stories from Lakon']. Collection of stories from the oral tradition of Lakon (Gaua, Banks Is). Monolingual in Lakon, for educational use. Illustrated by Sawako François. Alliance Française de Port-Vila. 44 pp.
- François, Alexandre. 2012. The media library project: Repatriating my audio archives. [Blog entry.] (<http://alex.francois.online.fr/AF-audio-library-e.htm>) (Accessed 2018-10-09).
- François, Alexandre. 2015. Literacy materials for the communities: Building up vernacular literacy. [Blog entry.] (<http://alex.francois.online.fr/AF-literacy-e.htm>) (Accessed 2018-10-09).
- François, Alexandre, Michael Franjeh; Sébastien Lacrampe & Stefan Schnell. 2015. The exceptional linguistic density of Vanuatu. In Alexandre François, Sébastien Lacrampe, Michael Franjeh & Stefan Schnell (eds.), *The Languages of Vanuatu: Unity and Diversity*. Studies in the Languages of Island Melanesia 5, 1-21. Canberra: Asia Pacific Linguistics Open Access.
- François, Alexandre & Edgar Howard. 2000. *Bulsal, dam galsi me lëklek* ['Follow me, my friend']. Language reader for vernacular education, monolingual in Mwotlap. Illustrated by Sawako François. 26 pp. Reprinted in 2006 with the help of 'Alliance Française' of Vanuatu.
- François, Alexandre & Monika Stern. 2013. *Musiques du Vanuatu: Fêtes et Mystères – Music of Vanuatu: Celebrations and Mysteries*. [Musical recordings] (Label INÉDIT). Paris: Maison des Cultures du Monde. CD album: 73'39", with booklet (128 pp.).
- Franjeh, Michael. 2018. *The languages of northern Ambrym, Vanuatu*. Archive of linguistic and cultural material from the North Ambrym and Fanbyak languages. London: SOAS, Endangered Languages Archive. (<https://elar.soas.ac.uk/Collection/MPI1143013>) (Accessed 2018-10-20).
- Frimigacci, Daniel, Muni Keletaona, Claire Moyses-Faurie & Bernard Vienne. 1995. *La tortue au dos moussue – Ko le fonu tu'a limulimua. Texte de tradition orale de Futuna*. Société d'Études Linguistiques et Anthropologiques de France 356. Louvain: Peeters.
- Green, Roger. 1991. Near and Remote Oceania: Disestablishing "Melanesia" in culture history. In Andrew Pawley (ed.), *Man and a half: Essays in Pacific Anthropology and Ethnobotany in honour of Ralph Bulmer*, 491-502. Auckland: The Polynesian Society.
- Hammarström, Harald, Sebastian Bank, Robert Forkel, & Martin Haspelmath. 2018. *Glottolog 3.2*. Jena: Max Planck Institute for the Science of Human History. (<http://glottolog.org>) (Accessed on 2018-10-18)

- Harris, Amanda; Nick Thieberger & Linda Barwick (eds.), *Research, records, and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures*. Sydney: University of Sydney Press. <http://hdl.handle.net/2123/13310>
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195.
- Hinton, Leanne, & Kenneth Hale (eds). 2001. *The Green Book of Language Revitalization in Practice*. San Diego, CA: Academic Press.
- Jacobson, Michel, Boyd Michailovsky, & John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication* 33.1-2 (2001): 79–96.
- King, Jeanette. 2001. Te Kōhanga Reo: Māori Language Revitalization. In Leanne Hinton & Ken Hale (eds.), *The Green Book of Language Revitalization in Practice* 118–128. San Diego, CA: Academic Press.
- Leenhardt, Maurice. 1946. *Langues et dialectes de l’Austro-Mélanésie*. Vol. 46. Paris: Institut d’ethnologie.
- Love, Jacob. 1979. Sāmoan variations. Cambridge, MA: Harvard University dissertation.
- Meyer, David. 2013. *Early Tahitian poetics*. (Pacific Linguistics 641.) Berlin: Walter de Gruyter.
- Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François & Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation & Conservation* 8. 119–135. <http://hdl.handle.net/10125/4621>
- Moyse-Faurie, Claire. 2012. Haméa et xârâgurè, langues kanak en danger. *UniverSOS: Revista de Lengüas Indígenas y Universos Culturales*, 73–86. Universitat de València.
- Nathan, David. 2010. Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing* 4(1-2). 111–124.
- Ozanne-Rivierre, Françoise. 1975–79. *Textes nemi, Nouvelle-Calédonie*. 2 vol. Paris: SELAF.
- Païa, Mirose. 2014. L’Enseignement des langues et de la culture polynésiennes à l’école primaire en Polynésie française. In Isabelle Nocus, Jacques Vernaudeau, & Mirose Paia (eds), *L’École plurilingue en Outre-mer: Apprendre plusieurs langues, plusieurs langues pour apprendre*, 409–429. Rennes: Presses Universitaires de Rennes.
- Palmer, Bill. 2018. Language families of the New Guinea area. In Bill Palmer (ed.), *The languages and linguistics of the New Guinea area: A comprehensive guide* (The World of Linguistics), 1–19. Berlin: de Gruyter.
- Paton, W.F. 1979. *Customs of Ambrym (texts, songs, games and drawings)* (Pacific Linguistics D-22). Canberra: Australian National University.
- Pawley, Andrew. 1981. Melanesian diversity and Polynesian homogeneity: a unified explanation for language. In Jim Hollyman & Andrew Pawley (eds.), *Studies in Pacific languages and cultures in honour of Bruce Biggs*, 259–310. Auckland: Linguistic Society of New Zealand.
- Pawley, Andrew. 2007. Why do Polynesian island groups have one language and Melanesian island groups have many? Patterns of interaction and diversification in the Austronesian colonization of Remote Oceania. Paper presented at the conference “Migrations”, September 5–7, 2007, Porquerolles, Var, France.
- Pawley, Andrew & Roger Green. 1973. Dating the dispersal of the Oceanic languages. *Oceanic Linguistics* 12(1/2): 1–67.

- Pearce, Katy E., & Ronald E. Rice. 2013. Digital divides from access to activities: Comparing mobile and personal computer Internet users. *Journal of Communication* 63(4). 721–744.
- Penfield, Susan D. & Benjamin V. Tucker. 2011. From documenting to revitalizing an endangered language: Where do applied linguists fit? *Language and Education* 25(4). 291–305.
- Ray, Sidney H. 1926. *A comparative study of the Melanesian island languages*. Cambridge: Cambridge University Press.
- Rivierre, Jean Claude, & Françoise Ozanne-Rivierre. 1980. *Mythes et contes de la Grande-Terre et des îles Loyauté (Nouvelle-Calédonie)*. Paris: SELAF.
- Simons, Gary & Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the world*, 21st edn. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com> (Accessed on 2018-10-18)
- Simons, Gary & Steven Bird. 2003. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing* 18(2). 117–128.
- Tcherkézoff, Serge. 2009. *Polynésie/Mélanésie: L'invention française des "races" et des régions de l'Océanie (xvi^e-xix^e siècles)*. Papeete: Au vent des îles.
- Thieberger, Nick & Linda Barwick. 2012. Keeping records of language diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). In Nicholas Evans & Marian Klamer (eds.), *Melanesian languages on the edge of Asia: Challenges for the 21st Century*, 239–253. *Language Documentation & Conservation Special Publication* 5. <http://hdl.handle.net/10125/4567>
- Tryon, Darrell T. 1976. *New Hebrides languages: An internal classification* (Pacific Linguistics C-50). Canberra: Australian National University.
- Turin, Mark, Claire Wheeler & Eleanor Wilkinson (eds.). 2013. *Oral literature in the digital age: Archiving orality and connecting with communities*. Cambridge: Open Book Publishers.
- Unseth, Peter & Lynn Landweer (eds.). 2012. *Language Use in Melanesia*. Special issue of *International Journal of the Sociology of Language* 214.
- Vanuatu Ministry of Education. 2012. *Vanuatu National Language Policy*. Official report, 20 pp.
- Vernaudon, Jacques. 2015. Linguistic Ideologies: Teaching Oceanic Languages in French Polynesia and New Caledonia. *The Contemporary Pacific* 27(2): 433–462. doi: 10.1353/cp.2015.0048
- Warner, Sam L. No'eau. 2001. The movement to revitalize Hawaiian language and culture. In Leanne Hinton & Ken Hale (eds.), *The Green Book of Language Revitalization in Practice* 133–144. San Diego, CA: Academic Press.
- Wittenburg, Peter, Ulrike Mosel & Arienne Dwyer. 2002. Methods of language documentation in the DOBES project. In Peter K. Austin, Helen Dry & Peter Wittenburg (eds.), *Proceedings of the international LREC workshop on resources and tools in field linguistics*, 36–42. Paris: European Language Resources Association (ELRA).

Alexandre François
alexandre.francois@ens.fr

 orcid.org/0000-0003-1947-0806

Reflections on language documentation in the Southern Cone

Fernando Zúñiga
University of Bern

Marisa Malvestitti
National University of Río Negro

Although many indigenous languages of Chile and Argentina have been documented only in the second half of the 20th century by academic anthropologists and linguists, some languages have a comparatively long tradition of descriptive and documentary scholarship conducted by Catholic missionaries. From a present-day perspective, early descriptions and documentations show some shortcomings (viz., they are often fragmentary and biased in several respects), but they nonetheless constitute a trove of valuable resources for later work and ongoing revitalization endeavors. Current documentary work is now more balanced in terms of Himmelmann's (1998) three-parameter typology (i.e., it pays close attention to communicative events of different kinds of modality, spontaneity, and naturalness), employs audio and video recordings, and takes copyright, access, and sustainability issues seriously. It is also more collaborative and empowering vis-à-vis the role played by indigenous collaborators than in the past and tends to be reasonably multidisciplinary.

Many indigenous languages of Latin America in general and of the Southern Cone in particular have not been documented until recently. The Chonan languages of Patagonia, for instance, drew the attention of linguists and anthropologists only in the 20th century. (Important forerunners include explorer-cum-chronicler Antonio Pigafetta in the 16th century, as well as physician-cum-anthropologist Robert Lehmann-Nitsche and several explorers in the 19th century; see Viegas Barros 2005.) Something similar happened regarding other languages of Chile (Kawésqar and Yahgan) and Argentina (Santiagoño Quechua and Selk'nam); for other languages of the Chaco region (Wichí, Toba, Mocoví, Pilagá, Chorote, Nivaclé, Tapiete, Kaiwá, Ava Guaraní, Vilela), see Golluscio & Vidal (this volume). Yet other languages were neglected and disappeared, leaving few traces outside

the translation of religious texts, onomastics, and a handful of everyday words, like the Charruan languages of Uruguay (Rosa 2013) and a dozen languages in Argentina and Chile. A notable exception is Diaguita-Calchaquí, also known as Cacán, documented as it was by Jesuit Alonso de Bárcena (1528–1598), but the manuscript is now lost.¹

Nevertheless, both language description and language documentation have a comparatively old history in Latin America: Catholic priests started describing and documenting indigenous languages—those few they saw as *lingua francas*—by the mid-16th century and continued working on them, albeit intermittently and irregularly, until the 20th century. (Like in other colonial contexts, the expansion of interethnic relations and territorial appropriation were accomplished with the support of linguistic records, whose accuracy and depth show significant variation.) In the Southern Cone, (first) Jesuits and (later) Capuchins authored full-fledged grammars and bilingual dictionaries, complemented by liturgical and catechistic texts in the early days and by collections of narratives in the early 20th century. Their aim was not only to enable missionaries to work in the areas where the languages were spoken but also to report on how considerable the development of such languages was as an intellectual achievement.

The most remarkable examples of this are Luis de Valdivia (1560–1642), who wrote complete but relatively brief descriptions of the northern variety of Mapudungun and the extinct Huarpean languages Allentiac and Millcayac, and Andrés Febrés (1734–1790), whose work on Mapudungun was either quoted from or integrated in different ways in almost every text dealing with that language in the 18th and 19th centuries (Malvestitti & Payás 2016).² Notably, missionary linguistics continued well after Spanish colonial rule had come to an end. In Chile, Félix José de Augusta (1860–1935) wrote a complete and fairly thorough description of the central variety of Mapudungun, and collected many texts (narratives and songs) from bilingual consultants who lived near the missions. In southern Patagonia and Tierra del Fuego, a number of Anglican priests from the South American Missionary Society authored language descriptions of Günün a Iajüch and especially Tehuelche and Yahgan,³ as did some Italian Salesians for Selk'nam and Kawésqar. Other missionaries helped them and continued their work (see, e.g., Moesbach 1930, 1962 and Molina 1967). Late missionary linguistics found intellectual support in the anthropological work conducted in the Southern Cone until the 1950s—which contrasted with the work conducted by professional linguists of the region until then, limited as it was to either European languages or historical-linguistic issues like genealogical relations and dialectal variation, as well as areal relationships and migration.⁴

This tradition was interrupted in the 1960s, when structural linguistics was introduced as the mainstream framework in Chile and Argentina. Centers for the study of indigenous languages were founded or further developed at several universities of the region, and their members started conducting linguistic fieldwork anew, with new theoretical

¹Even though Easter Island is part of one of the Chilean territorial administrative units, we do not address its Polynesian language (Rapa Nui) here, because it belongs to the Pacific rather than to the Southern Cone.

²As far as the Andean Plateau is concerned, several early descriptions are worthy of mention: Domingo de Santo Tomás's (1499–1570) of Classical Quechua, Diego de Torres Rubio's (1547–1538) of Bolivian Quechua and Aymara, Ludovico Bertonio's (1557–1625) of Aymara, and Diego González Holguín's (1560–1620) of Cuzco Quechua. In addition, Antonio Ruiz de Montoya (1585–1652) authored the first description of Paraguayan Guaraní and early descriptions of several languages of Mexico are also well known and important; see Brevia-Claramonte (2007, 2008).

³*Günün a iajüch* is the endonym of the language also known as Puelche and Northern Tehuelche (ISO-code: pue). Tehuelche is also known as Aonikenk and Southern Tehuelche (endonym: *aonek'o 'a'jen*; ISO-code: teh).

⁴See, e.g., the valuable Mapudungun text collections by Lenz (1895–1897) and by Lehmann-Nitsche (Malvestitti 2012).

foundations and methods (including audio recordings). At first, some languages were documented rather sketchily, via texts and word lists (e.g., Tehuelche and Selk'nam, by Jorge Suárez 1966–1968, n.d., and Mapudungun, by Golbert 1975). They were described more comprehensively later, via full-fledged dictionaries, grammars, and in-depth studies (e.g., Selk'nam by Najlis 1973; Yahgan by Golbert 1977, 1978; Kawésqar, by Clairis 1987 and Aguilera 2000; Tehuelche, by Fernández Garay 1997, 1998; Mapudungun, by Sánchez 1989 and Salas 1992,⁵ among others). By contrast, very few studies were conducted from the perspective of ethnography of communication (e.g., Golluscio 2006).

Early work by missionaries is both the basis for subsequent work and a valuable depository of linguistic, anthropological, and historical information on several indigenous societies of the region. By present-day standards, however, such work shows some shortcomings. Rather than describing the languages for the benefit of academic disciplines, the main objective was to assist learning by L2 speakers involved in missionary endeavors, which led to some domains of language structure and use being inadequately covered and other domains being ignored. The missionaries did not work haphazardly and were acquainted with contemporary British, Italian, and German philological and anthropological scholarship. Nevertheless, the blueprint employed was centered on Nebrija's (2011 [1492]) Spanish grammar and didactic grammars of Latin and Ancient Greek, and the treatment given to local customs and religious issues regularly betrayed a prejudiced attitude on the part of the authors.⁶

Moreover, documentations were fragmentary and skewed: until the late 20th century, they registered communicative events only exceptionally, and without providing any metalinguistic information. In terms of Himmelmann's (1998) three-parameter typology of communicative events (i.e., modality, spontaneity, and naturalness), even those texts collected since the 1960s recorded almost exclusively oral texts and clearly favored those on the planned and less natural ends (viz., interviews, narratives, monologues, ritual speeches, and elicitation). Language documentation in the Southern Cone transitioned from paper-based formats to those including audio and video recordings by the beginning of the 21st century.⁷

Present-day scholars regard locating, retrieving, and digitizing early written materials that are difficult to obtain as an important task in its own right. Not only does such "declassification" (Pavez Ojeda 2008) of sources preserved in numerous and disparate private and institutional sites help to reassess the proficiency of some language consultants and to contextualize the documentation practices developed in the region. It can also supply new valuable data recorded at a time when the languages were still in everyday use that were simply not considered worthy of publication then.⁸ This reconnecting of field data and the situations in which they originated with the present-day language communities is particularly important in the case of languages hitherto regarded as terminally endangered or even extinct, like Günün a Iajüch and Tehuelche:

⁵Adalberto Salas brought a fresh approach to the analysis of Mapudungun grammar and greatly contributed to the development of Mapudungun studies in Chile from the 1970s until his death in 2000.

⁶There is a sizable volume of late-20th-century and early-21st-century literature on how to properly contextualize and assess strengths and weaknesses of early language descriptions in Mesoamerica and South America; see, e.g., Brevia-Claramonte (2007, 2008, 2009) and the references therein.

⁷Since the early 20th century, the intention to make audio recordings was mentioned in the linguistic reports, and from the 1950s onwards recording technologies were used in fieldwork. Nevertheless, the publication of audio recordings by linguists was unusual; see, as noteworthy exception, Fernández (1985).

⁸Notable examples of this include information on the composition strategies and transliteration issues in the drafts written by Mapuche collaborators of some well-known scholars, viz. Mankilef for Tomás Guevara (Pavez Ojeda 2003) and Nahuelpi for Lehmann-Nitsche.

it provides valuable support to ongoing initiatives of language revitalization conducted by neo-speakers with the collaboration of teachers and linguists. In this context, it is fitting to mention repositories like the Laboratorio de Documentación e Investigación en Lingüística y Antropología (DILA) in Argentina (which depends on the Consejo Nacional de Investigaciones Científicas y Técnicas) and the Centro de Documentación Indígena (CDI) and Aike Biblioteca Digital de la Patagonia in Chile, as well as the Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin.⁹

As far as the limits of documentation are concerned, missionaries and anthropologists were not alone in neglecting the ethical issues that present-day scholars, academic institutions, and funding agencies aptly take so seriously. The bulk of the indigenous-related public policies implemented by the Chilean and Argentinian states during the 19th century and most of the 20th century ranged from racist to assimilationist and paternalistic; cultural and educational institutions have come to safeguard some of the interests of indigenous communities in a systematic fashion only recently. Despite scholarly and legal concerns with intellectual property and the dissemination of potentially sensitive information, field linguists working in the region sporadically face the situations that their fellow researchers routinely report for North America and Australia. Language consultants willing to work with linguists and anthropologists vary greatly with respect to their level of (Western) education, their occupation and place of residence, and their involvement in, and support of, community-oriented political or cultural activities. Such heterogeneity notwithstanding, consultants have only recently started to ask researchers to handle issues related to copyright and access to material in a particularly restrictive way. (Such safety measures have normally been imposed by funding agencies and ethics committees since the early 1990s.) To our knowledge, no indigenous society in the Southern Cone shows a picture like the one described for the Rio Grande Pueblos by Brandt (1980, 1981) and mentioned by Himmelmann (1998), where language-mediated religious and ceremonial knowledge is intimately related to political and cultural leadership in such a way that language documentation is inevitably and significantly disruptive.

Unlike in other parts of the world, scholarly activity in the Southern Cone is typically not regarded as exploitative by the indigenous communities. To be sure, interaction with the dominant society has disrupted the communities' traditional way of life and compromised their viability, which has often eclipsed the fact that their worldview has been called into question and their language endangered. To the extent that an ideology of conflict has been explicitly formulated at all in recent times, however—which is perhaps most evidently the case with the customarily belligerent Mapuche—, linguists and anthropologists are usually seen as helpful, and often friendly, intermediaries between the indigenous and the non-indigenous societies.¹⁰ Among other things, scholars provide advice and support endeavors like literacy training and revitalization efforts, as coaches, trainers, and/or fund raisers. Even though sustainable community-based initiatives in which non-indigenous scholars only play a secondary role are still difficult to implement,

⁹We are grateful to reviewer for pointing out to us that Anthony Woodbury, in representation of the AILLA, deposited in 2009 a copy of all recordings of indigenous-language materials collected by Argentinian researchers existing in AILLA in the DILA-CONICET Archive, including Tehuelche recordings by Jorge Suárez and Emma Gregores and Mapudungun recordings by Lucía Golluscio.

¹⁰This is not a recent phenomenon. For instance, Lehmann-Nitsche's consultants called him *dear doctor* or *inche ñi kume ueni* 'my good friend' (Malvestitti 2012: 55), and Chapman (2002) highlights the friendship established with her main consultants. References of interlocutors as *teacher* and *collaborator* often appear in diverse early-20th-century sources.

especially in the case of small groups with very limited resources, research is conducted in a less paternalistic and more empowering manner than three or four decades ago.¹¹

In fact, scholars belonging to the largest indigenous language community of the region started to become active and visible at about that very time. Working outside of academia, Chilean linguist and historian Armando Raguileo (1922–1992) proposed an alternative orthography for the writing of Mapudungun in the mid-1980s (i.e., the so-called *grafemario Raguileo*), which was largely adopted (and later slightly adapted) by non-linguists on both sides of the Andes. Since the late 1980s, teacher Segundo Llamín Canulaf (1926–) and other Mapuche authors have written and published their own educational and historical bilingual texts in Chile, and several Argentinian Mapuche have authored descriptive studies, as well as literary texts (e.g., Ministerio de Educación 2015, Equipo de Educación Mapuche Wixaleiñ 2015). Within Chilean academia, Mapuche scholars like María Catrileo since the 1980s and Elisa Loncon since the late 2000s have significantly contributed to Mapudungun studies with a focus on education and revitalization.¹²

Finally, Himmelmann's article pertinently emphasizes the importance of a language documentation that does not cater exclusively, or even primarily, to linguistic typologists and theoreticians. Analytically inadequate and culturally skewed though it was, pre-modern language documentation in the Southern Cone did leave a lasting legacy of multidisciplinary—rather unsurprisingly so, rooted as it was in a pre-disciplinary approach employed by non-professional practitioners. Present-day academia strives to counteract some of the negative consequences of ever-deepening specialization in increasingly fragmented disciplines by fostering interdisciplinarity (where individual disciplinary approaches are not only contrasted but also integrated) and transdisciplinarity (where individual disciplines ideally dissolve into novel holistic approaches). Nevertheless, descriptive linguistics in the Southern Cone has barely finished consolidating its status as a discipline in its own right—much along the lines described by Himmelmann (1998). To judge not only from which documentation projects compete, both locally and internationally, for institutional funds and academic validation but also from how successful projects are conducted nowadays, the most healthful and promising pressure to develop some actual interdisciplinarity stems from collaboration with anthropology and educational endeavors.

¹¹Even early linguistic research conducted in Patagonia, for example, routinely acknowledged the work of particular (elderly) consultants that had a good memory and were especially talented performers of verbal art (e.g., Borgatello 1928 and Harrington 1946).


¹²See Aguilera & Tonko (2009) for Kawésqar. For almost every language of the region, a growing number of written and multimedia productions developed by indigenous collectives intend to link previous documentations with ongoing revitalization endeavors.

References

- Aguilera, Óscar. 2000. *Kawésqar*. Munich: Lincom Europa.
- Aguilera, Óscar & José Tonko. 2009. *Cuentos kawésqar*. Santiago: FUCOA, Ministerio de Agricultura.
- Borgatello, Maggiorino. 1928. *Notizie grammaticali e glossario della lingua degli indii alakaluf abitanti dei canali magellánicos della Terra del Fuoco*. Torino: Società Editrice Internazionale.
- Brandt, Elizabeth. 1980. On secrecy and control of knowledge. In Tefft, Stanton (ed.), *Secrecy: A cross-cultural perspective*, 123–146. New York: Human Sciences Press.
- Brandt, Elizabeth. 1981. Native American attitudes toward literacy and recording in the Southwest. *Journal of the Linguistic Association of the Southwest* 4. 185–195.
- Breva-Claramonte, Manuel. 2007. The European linguistic tradition and early missionary grammars in Central and South America. In Douglas Kibbee (ed.), *History of linguistics 2005: Selected papers from the Tenth International Conference on the History of the Language Sciences, 1–5 September 2005, Urbana-Champaign, Illinois*, 236–251. Amsterdam: John Benjamins.
- Breva-Claramonte, Manuel. 2008. El marco doctrinal de la tradición lingüística europea y los primeros misioneros de la Colonia. *Bulletin Hispanique* 110(1). 25–59.
- Breva-Claramonte, Manuel. 2009. *La didáctica de las lenguas en el Renacimiento: Juan Luis Vives y Pedro Simón Abril, con selección de textos*. Bilbao: Universidad de Deusto.
- Chapman, Anne. 2002. Introducción. In Anne Chapman (ed.), *Fin de un mundo: los selknam de Tierra del Fuego*, 12–18. Santiago: Taller Experimental Cuerpos Pintados.
- Clairis, Christos. 1987. *El qawasqar. Lingüística fueguina, teoría y descripción*. Valdivia: Universidad Austral de Chile.
- Equipo de Educación Mapuche Wixaleñ. 2015. *Diccionario básico de idioma mapuche*. Florencio Varela: Xalkan Ediciones.
- Fernández, César. 1985. *Romanceadas mapuches*. Cassette tape edition. Neuquén: Facultad de Humanidades, Universidad Nacional del Comahue.
- Fernández Garay, Ana. 1997. *Testimonios de los últimos tehuelches: textos originales con traducción y notas lingüístico-etnográficas*. Buenos Aires: Instituto de Lingüística, Universidad de Buenos Aires.
- Fernández Garay, Ana. 1998. *El tehuelche: una lengua en vías de extinción*. Valdivia: Universidad Austral de Chile.
- Golbert, Perla. 1975. *Epu peñiwen “Los dos hermanos”: cuento tradicional mapuche*. Buenos Aires: C.I.C.E.
- Golbert, Perla. 1977. Yaghán I: las partes de la oración. *VICUS Cuadernos Lingüística* 1: 5–60.
- Golbert, Perla. 1978. Yaghán II: morfología verbal. *VICUS Cuadernos Lingüística* 2: 87–101.
- Golluscio, Lucía. 2006. *El pueblo mapuche: poéticas de pertenencia y devenir*. Buenos Aires: Biblos.
- Golluscio, Lucía & Vidal, Alejandra. 2018. Reflections on language documentation in the Chaco. In Bradley McDonnell, Andrea Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation on the 20 year anniversary of Himmelmann 1998*, 303–320. (Language Documentation & Conservation Special Publication 15.) Honolulu: University of Hawai‘i Press. <http://hdl.handle.net/10125/24831>

- Harrington, Tomás. 1946. Contribución al estudio del indio Gününa Küne. *Revista del Museo de La Plata* II(14). 237–275.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195.
- Lehmann-Nitsche, Roberto. 1915–1916. *Vocabulario Puelche*. Berlin: Ibero-Amerikanisches Institut.
- Lenz, Rodolfo. 1895–1897. *Estudios araucanos*. Santiago: Imprenta Cervantes.
- Malvestitti, Marisa. 2012. *Mongeluchi zungu: los textos araucanos documentados por Roberto Lehmann-Nitsche*. Berlin: Ibero-Amerikanisches Institut / Gebr. Mann Verlag.
- Malvestitti, Marisa & Gertrudis Payás. 2016. Circulaciones intertextuales del *Arte* de Febrés a ambos lados de los Andes. In María Andrea Nicoletti & Paula Núñez (eds.), *Araucanía-Norpatagonia: discursos y representaciones de la materialidad*, 305–331. Viedma: Editorial de la Universidad Nacional de Río Negro; San Carlos de Bariloche: Instituto de Investigaciones en Diversidad Cultural y Procesos de Cambio.
- Ministerio de Educación. 2015. *Con nuestra voz estamos: Escritos plurilingües de docentes, alumnos, miembros de pueblos originarios y hablantes de lenguas indígenas*. Buenos Aires: Ministerio de Educación de la Nación.
- Moesbach, Ernesto Wilhelm de. 1930. *Vida y costumbres de los indígenas araucanos en la segunda mitad del siglo XIX*. Santiago: Imprenta Cervantes. (Later reissued as Coña, Pascual. 1973. *Memorias de un cacique mapuche*. Santiago: ICIRA.)
- Moesbach, Ernesto Wilhelm de. 1962. *Idioma mapuche*. Padre Las Casas: Imprenta San Francisco.
- Molina, Manuel. 1967. Antiguos pueblos patagónicos y pampeanos a través de las crónicas. 1ª y 2ª parte. *Anales de la Universidad de la Patagonia "San Juan Bosco"* 3(I). 19–184.
- Najlis, Elena. 1973. *Lengua selknam*. Buenos Aires: Instituto de Filología y Lingüística, Facultad de Historia y Letras, Universidad del Salvador.
- Nebrija, Antonio de. 2011 [1492]. *Gramática sobre la lengua castellana*. (Ed. Carmen Lozano.) Madrid: Real Academia Española / Galaxia Gutenberg.
- Pavez Ojeda, Jorge. 2003. Mapuche ñi nütram chillkatun / Escribir la historia mapuche. Estudio posliminar de "Trokinche mufu ñi piel". *Historias de familias siglo XIX. Revista de Historia Indígena* 7. 7–53.
- Pavez Ojeda, Jorge. 2008. *Cartas mapuche: siglo XIX*. Santiago: CoLibris / Ocho Libros.
- Rosa, Juan Justino da. 2013. Historiografía lingüística del Río de la Plata: las lenguas indígenas de la Banda Oriental. *Boletín de Filología* XLVIII(2). 131–171.
- Salas, Adalberto. 1992. *El mapuche o araucano*. Madrid: MAPFRE.
- Sánchez, Gilberto. 1989. Relatos orales en pewenche chileno. *Anales de la Universidad de Chile, Estudios en Honor de Yolando Pino Saavedra, Quinta Serie*, 17. 289–360.
- Suárez, Jorge. 1966–1968. Lista de palabras y frases en Tehuelche: Proyecto Tehuelche. (<http://www.ailla.org>) Archivo de Lenguas Indígenas de Latinoamérica. TEH001R001–TEH001R020.
- Suárez, Jorge. n.d. Lista de palabras en Ona. Proyecto Ona. (<http://www.ailla.org>) Archivo de Lenguas Indígenas de Latinoamérica. ONA001R001–ONA001R005.

Viegas Barros, José Pedro. 2005. *Voces en el viento: raíces indígenas de la Patagonia*. Buenos Aires: Mondragón.

Fernando Zúñiga
fernando.zuniga@isw.unibe.ch
 orcid.org/0000-0002-9015-6601

Marisa Malvestitti
mmalvestitti@unrn.edu.ar
 orcid.org/0000-0002-0798-8408

Reflections on language documentation in the Chaco

Lucía Golluscio

*Consejo Nacional de Investigaciones Científicas y Técnicas
Universidad de Buenos Aires*

Alejandra Vidal

*Consejo Nacional de Investigaciones Científicas y Técnicas
Universidad Nacional de Formosa*

This chapter focuses on field research aimed at documenting Chaco languages with varying degrees of vitality, specifically those spoken in Argentina and in the vicinity of the Argentinian/Paraguayan, Argentinian/Bolivian, and Paraguayan/Bolivian borders. The case studies here selected provide an overview of recent experiences conducted in Chaco within the framework of Himmelmann 1998's foundational program on documentary linguistics and subsequent publications along these lines. We emphasize the results of collaborative research on equal grounds and a discourse-oriented approach to language documentation. Our reflections also highlight the current threatening situation of indigenous peoples and their languages and discuss the function of language documentation, preservation, and archiving in this fragile scenario, with a view to supporting community language use and transmission as well as ongoing and future research in South America.

1. Introduction¹ In the early 1990s, a relevant publication in *Language* by Hale et al. (1992), placed the topic of endangered languages around the world on the international academic agenda. Explicitly linked to this topic, the seminal paper by Himmelmann (1998) brings the issue of linguistic documentation to the forefront while advocating for documentary linguistics with an autonomous status similar to descriptive linguistics.

¹We are very grateful to the editors of LD&C for their invitation to collaborate in this volume and their recommendations to our first manuscript as well as to two anonymous reviewers for their comments and suggestions. Our special thanks go to the following colleagues for their generous contribution to the article: Elizabeth Birks, Florencia Ciccone, Luca Ciucci, Santiago Durante, Hebe González, Analía Gutiérrez, and Verónica Nercesian.

In this article, we focus our reflection on contemporary collaborative research projects on Chacoan languages developed within the documentary linguistics framework. After this introduction (§1), in §2 we briefly consider a set of case studies; then, based on the results of those initiatives, we reflect on language documentation in the region and in South America (§3). Finally, in §4 we present the proposal for the creation of the South American Network of Regional Linguistic and Sociocultural Archives, with a view to the present and future work on language documentation and preservation, including the issue of data accessibility and exchange.

Chaco, in the heart of South America, comprises a vast lowland territory ranging from southeast Bolivia and the southwestern area of the Mato Grosso in Brazil northwards, to the westernmost area of Paraguay and northeast of Argentina southwards. It is a plurilingual region where twenty languages belonging to seven linguistic families are spoken with differing degrees of vitality. Moreover, extended multilingualism phenomena have been registered on the Bolivia/Argentina/Paraguay border (Campbell & Grondona 2010; Ciccone 2015).²

The panorama among Chaco peoples is particularly complex. Many of them inhabit lands currently belonging to different countries and, therefore, ruled by different socio-educational policies. In Argentina, Intercultural Bilingual Education was until recently a program under the National Education Law enacted in 2006. However, resistance from many sectors, including some of the teachers themselves (Vidal & Kuchenbrandt 2015: 91), and the limited availability of materials for bilingual education mean these peoples are not offered equal opportunities within the educational system and literacy in their languages is not always valued by national or regional governments.

The situation regarding linguistic vitality is also heterogeneous. On the one hand, there are communities where use and transmission of the heritage language is supported by collaborative experiences in documentation and ongoing linguistic description, with community-led linguistic activism (2.2, 2.3, 2.4). On the other, important language attrition and shift processes have been documented (2.1). In particular, our field research has shown that transmission to younger generations does not always occur (2.1, 2.4). Language transmission and use is further affected by migration to the cities, where native language teaching is seldom on the curriculum. Living in rural settings helps strengthen shared socio-cultural ties. However, after moving to the cities, speakers usually become a minority within a Spanish-speaking majority. See the following striking comparative data (INDEC-ECPI, 2004-2005). Whereas less than half of Pilagá (48%) and Wichí (35%) speakers live in urban areas in Argentina, the Toba/Qom proportion of urban population is much higher (69%). This Census reports 99% and 91% of Pilagá and Wichí native speakers, respectively, while the Toba/Qom speaker proportion is 65%. In spite of these figures, a tendency towards incomplete language acquisition has been observed among Pilagá (2.4); see also Tapiete (2.1). Hence, speakers of these languages in peri-urban settlements do not achieve skills in some registers of their heritage language (principally, oratory and narrative).

In Argentina, these processes have been triggered within a broader context defined by a hegemonic Hispanizing language ideology fostered by the national state since the second half of the 19th century. Other factors include speakers' marginal status; changes in organization, from hunter-gatherer to (semi)urban sedentary lifestyle; migration and

²For a complete view of the distribution and situation of peoples and languages in the area, see Unruh & Kalisch (2003), Censabella (2009), Lewis (2009). Detailed references on Chaco languages studies can be found in Fabre (1998; 2017a [2005]), Golluscio & Vidal (2009-2010), and Campbell & Grondona (2012).

integration with other indigenous or non-indigenous communities; everyday contact with the Spanish-speaking population and media, and a diminished sense of ethnic pride caused by racism and discrimination. In some cases, governments' denial of these peoples' Constitutional rights contributes to the situation. Even for those currently settled in rural enclaves in Argentina, pressure by the spreading agricultural frontier into the forested territories they inhabit is a crucial problem that puts not only their culture and language transmission, but also their very survival at risk, not to mention serious local and global damage caused by deforestation and soy plantations.³

Fortunately, the communities here considered have shown a great deal of interest in language revitalization. The still everyday use of their languages ensures secure format documentation to produce linguistic materials for educational purposes. These are the reasons that have led to undertaking linguistic documentation of most Chaco languages over the last twenty years (see §2).

2. Documentation of Chaco Languages focusing on Argentina and Paraguay

Fieldwork-based linguistic research by professional linguists focusing on Chaco languages began in the 1960s and 1970s, and included grammars, vocabularies, and phonological studies. From a descriptive perspective, they provided an analysis and written documentation of these languages. Subsequently, work on Chaco languages increased, especially through doctoral dissertations. Though relevant to our past and current knowledge of the Chaco, these studies were not necessarily intended for language documentation, neither did they adopt a discourse-oriented approach.

It was not until the early 21st century that practices and research projects within the framework of documentary linguistics on Chaco languages began. The first was the Chaco languages Project (2002-2005), "Endangered Languages, Endangered Peoples in Argentina. Documentation of four Chaco languages in their ethnographic context: Mocoví, Tapiete, Vilela, and Wichí", carried out as part of the Dokumentation Bedrohter Sprachen (DoBeS) Program, under the auspices of the Volkswagen Foundation (see Golluscio & Hirsch 2006).⁴ Audio and video resources from this project, mostly having open access, have been deposited both in the DoBeS Archive and the Archive of the Laboratorio de Documentación e Investigación en Lingüística y Antropología (DILA) (<http://www.caicyt-conicet.gov.ar/dila>), created at the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, by agreement with the Max Planck Institute (MPI) for Psycholinguistics, Nijmegen, The Netherlands, in 2007.

In the last two decades, there has been a significant increase in available audio and video records of Pilagá, Wichí, Chorote, Nivaêle, Mocoví, and Ayoreo language incorporated to the Endangered Languages Archive (ELAR) within the framework of the Endangered Languages Documentation Programme (ELDP) and the Hans Rausing Endangered Languages Project (HRELP). Likewise, Toba/Qom, Maká, and Mocoví materials can be found in the Archive of the Indigenous Languages of Latin America (AILLA), which, in 2009, donated a copy of these resources to DILA-CONICET.

³For more information on the socio-political and sociolinguistic situation in Argentina and Chile, see Zúñiga & Malvestitti, this volume.

⁴This collaborative documentation project settled at the University of Buenos Aires (UBA) was conducted by an interdisciplinary team under Lucía Golluscio's supervision in academic collaboration with Bernard Comrie (Department of Linguistics, MPI for Evolutionary Anthropology) (<http://www.mpi.nl/DOBES/projects/chaco>).

Recent experiences, fieldwork techniques and results from some of the Chaco language documentation projects will be reviewed in the following subsections. We selected projects focused on languages belonging to different families with varying degrees of vitality and developed under international programs such as DoBeS, HRELP, the Foundation for Endangered Languages (FEL), and the Documenting Endangered Languages Program (DEL) from the National Science Foundation (NSF). These projects taken together represent a paradigmatic change in the field of Chaco linguistics, given the collaborative perspective and the technological possibilities for disseminating results. For many researchers in Argentinean academia, the focus on language work has shifted to include recording discourse exchanges and verbal art –see Messineo (2008) in this journal, among others– and set the records in formats useful to a number of parties, from communities and linguists to social organizations and national and provincial education ministries.

2.1 Tapiete and Vilela, two endangered languages in the Argentine Chaco Tapiete (Tupí-Guaraní) and Vilela (Lule-Vilela, affiliation under discussion) are the two most endangered languages in the Argentine Chaco. Both show evidence of linguistic attrition without obsolescence. However, their current differing sociolinguistic and sociopolitical situations raise very different opportunities for their future (Golluscio & González 2008). While multigenerational, closely-knit Tapiete communities do exist and have a relatively small number of speakers (about 2,200, Ciccone 2015) distributed around Argentina, Paraguay and Bolivia, the extreme paucity of Vilela speakers (only two speakers, actually remembers, have been located) and the lack of a speech community have proven to be critical threats for this language. Moreover, the interruption of Tapiete language transmission is quite recent (González 2005). In contrast, the current Vilela situation is the result of a lengthy cultural and political de-structuring process intensified in the second half of the twentieth century. The elderly speakers who are working collaboratively on the documentation of their language (ML and GC, siblings, now 85 and 83), though exposed to Vilela on a daily basis during their childhood and youth, do not currently use it in their everyday life (Domínguez et al. 2006).

The existence of a Tapiete speech community enabled researchers to adopt traditional fieldwork techniques with older speakers, but linguistic tasks involved in documenting contemporary ways of speaking among younger speakers proved to be methodologically challenging (Ciccone 2015). Younger bilingual speakers incorporate loanwords, grammatical innovations and extensive code-switching, without participating in the performance of traditional narratives. It was necessary to foster particular situations to elicit spontaneous performances in Tapiete (see Figure 1). Traditional linguistic techniques would not have enabled the recording of verbal exchanges among these younger speakers, given their high regard for expert Tapiete speakers and their recognized abilities as language consultants, which the younger ones feel they cannot match.

As part of documenting traditional knowledge, an ethnobotanic dictionary is underway, in collaboration between Hebe González and members of the Tapiete communities in Salta, Argentina (González 2017 and in press) (see Figure 2). Apart from its contribution to lexical and cultural studies, this work includes a collection of analyzed narratives on Tapiete life in the Chaco.



Figure 1: Tapiete people. Awara Montes, Florencia Ciccone and members of the community. Tartagal, Salta, Argentina.



Figure 2: Helena Cabeza (Tapiete community) holding a specimen of *ñambi* ‘spicy herb’ (*Acmella oppositifolia*). Photo by Hebe González.

Vilela documentation has, instead, been based on a single-speaker-centered approach, including both the documentation of linguistic attrition in our main consultant's speech and the systematization of language-remembering processes triggered during his participation in documenting his language. The remembering strategies achieving best results (see Figures 3 and 4) were his return home and joint work with his sister. Narratives about topics relating to their own childhood, when the mother tongue becomes engrained, were key to motivating remembering of the language. This situation brings up an interesting question on the existence of latent cognitive strata that may resurface when stimulated by emotionally-charged experiences. Finally, collaborative linguistic research highlights the essential contribution of these last speakers' generation to the knowledge of: their heritage language, evidences of contact with other languages in and beyond the area (Golluscio 2015), and the persistence of Vilela structural characteristics and lexicon since the 18th century (Zamponi & Golluscio 2018).



Figure 3: Vilela language documentation. Returning to his place of origin. Mario López with Analía Gutiérrez, team member. Photo by Marcelo Domínguez.



Figure 4: Recovering Vilela basic vocabulary. Art session at Mario López' house with his grandchildren, great-nephews and María Hellemeier, team member. Photos by L. Golluscio.

2.2 Ayoreo discourse-oriented documentation The second project involves collaborative documentation of Ayoreo discourse conducted in Campo Loro, Paraguay. Ayoreo (Zamucoan) is still the language of communication in Paraguayan and Bolivian Ayoreo communities, although some signs of linguistic attrition have been documented (Durante p.c.). The 2012 census data for the above-mentioned countries estimate 1,862 Ayoreo people in Bolivia (CEDIB 2012) and 2,481 in Paraguay (DGEEC 2012), with some communities as yet uncontacted in Paraguay. The speech community exhibits speakers of all ages with the elderly and youngsters being mostly monolingual. The current vitality of the language is clear in the narratives collected in close collaboration with the community by Santiago Durante, a former PhD student at the University of Buenos Aires, under the auspices of ELDP (<https://elar.soas.ac.uk/Collection/MPI192274>). However, the presence of Spanish at school, in the media and in the social networks is jeopardizing the future of the language.

This collaborative research has centered on the documentation of texts of various genres in high-quality audio and video recordings. The collected stories, about life before contact, evangelization and sedentarization, are of great interest, since Campo Loro inhabitants have only recently come into contact with non-indigenous groups. The outcomes of the analyzed and annotated text-corpus include a significant volume of new information on Ayoreo grammatical structures based on naturalistic data and the publication of a collaborative anthology of narratives (Etacore & Durante 2016) already in use in the community. The book was welcomed by community members who interpret it not as a finished product but as a first step in the process of documenting their cultural and linguistic heritage (see Figure 5).



Figure 5: Documenting Ayoreo. Benito Etacore and Santiago Durante editing the book *Campo Loro gosode oe ojñane udojo*, Boquerón Department, Paraguay.

⁴Our knowledge of the Zamucoan languages (Ayoreo and Chamacoco) owes much to Pier Marco Bertinetto and Lucca Ciucci (Figure 6). See complete references on these authors in Fabre (2017b).



Figure 6: Documenting Chamacoco. Luca Ciucci with Francisco García and Domingo Calonga, Paraguay.

2.3 Nivaçle: a single language and territory, two countries There are an estimated 1,000 Nivaçle speakers in the provinces of Salta and Formosa, Argentina, and 12,000 in Paraguay. Today, there are open access audio resources for the community and more broadly for linguists (<https://elar.soas.ac.uk>). With support from ELDP and CONICET, Analía Gutiérrez has been investigating dialectal differences in Paraguay and contributing to capacity-building for language transcription and decision-making with regard to competing alphabets (Gutiérrez 2015) (see Figure 7). In Paraguay, this language is the primary means of communication among family members within indigenous communities, but Spanish and to some extent Guaraní are used with outsiders (Fabre 2017b). There are incipient bilingual programs in Paraguayan Nivaçle community schools. In Salta, the community is multilingual, living in peri-urban settings. Nivaçle communities in Formosa have received no attention from the local government and there are no bilingual education programs to serve around 93 school and pre-school-aged children from 180 families. Communities are denied most civil rights and recognition as an indigenous group. There is extended Nivaçle-Spanish bilingualism in Formosa but no extended multilingualism as documented in the Bolivia/Argentina/Paraguay border (see §1). In recent years, a project awarded by NSF to Alejandra Vidal as co-principal researcher (see 2.4) has enabled the production of audio and video resources in Nivaçle spoken in Formosa, Argentina. Transcription, analysis and translation of 5 hours of texts (conversations, narratives, songs) for archiving is currently underway. A selection of Nivaçle narratives was published for community use (Vidal 2015).



Figure 7: Work session at the First Meeting of Nivaçle Teachers, Uj'e Lhavos, Paraguay, Photo by Analía Gutiérrez.

2.4 Collaborative documentation, description and revitalization activities of Wichí and Pilagá Wichí language, spoken across the borders of Argentina and Bolivia, is transmitted intergenerationally. Characterized by their visibility, with their own radio programs and political organizations, representatives at government levels, some primary and secondary school teachers, the Wichí are one of the most numerous groups in the region (around 40,000 people).

Human resources training was a central issue in the DoBeS Chaco project (see page 295) and continues to be so. Linguistic training of younger Wichí speakers from settlements located by the Bermejo River, collecting oral discourse, developing reading materials, vocabularies and grammar courses for the study of their language, culture and history were activities pursued by Verónica Nercesian (Nercesian 2014a) (see Figure 8). Currently, the study of Wichí/Weenhayek dialectal varieties and sociohistorical processes in Northern Argentina and Southern Bolivia is underway. A published Wichí grammar (Nercesian 2014b) shows significant progress in new lines of research, such as the interaction between phonology, morphology, syntax and semantics. This model has opened up a new perspective on the study of similar phenomena in other Chaco languages.



Figure 8: Recording session for the Oral History Archive, Ramón Lista, Formosa, Argentina. Photo by Verónica Nercesian.

Pilagá, with its 5,000 speakers, does not enjoy the vitality it did 20 years ago when Vidal's research on the language began (Vidal 2001). Although intergenerational transmission of the mother tongue still occurs in some rural communities, and language teaching materials such as a talking dictionary and a learners' grammar have been developed for community schools (funded by ELDP and FEL; see Vidal & Miranda 2010;

Vidal, Almeida & Miranda 2014a–d), there are symptoms of linguistic attrition especially in semi-urban and urban settlements.

A recent documentation project (NSF-DEL 263817) was established between the University of Oregon (with Doris Payne as PI) and the Universidad Nacional de Formosa, Argentina, with the purpose of obtaining high-quality audio and video recordings in Pilagá and Nivaçle spoken in Formosa, and text-collection for archiving (see also 2.3). These include narratives on past conflicts and contact between Pilagá, Nivaçle and Wichí groups (see Figure 9). The texts support very fragmentary data provided by earlier European travelers and ethnographers about encounters with indigenous peoples in the Chaco region and their distribution in the territory where fieldwork is conducted.



Figure 9: Alejandra Vidal recording Pilagá stories, Km 30, Formosa, Argentina.

3. Directions in Chaco language documentation: Thoughts from the field The initiatives on language documentation presented in §2 raise a number of issues. Following Himmelmann's design for a language documentation project, our considerations evolve in three directions: corpus collection, corpus theorization and the role of the participants.

Regarding *corpus collection*, the fragile situation and degree of endangerment of Chaco languages described above provide a strong warning that, although linguistic analyses may be conducted at a later date, documentation of indigenous languages in Chaco and South America is extremely urgent with much work still pending. On the other hand, documentation of these languages should allow for developing versatile fieldwork methodologies to deal with a range of speakers in different situations and scenarios.

As to *corpus theorization*, we believe that the sharp distinction between documentation and description proposed in Himmelmann (1998) had a foundational epistemological function and relevance: It was necessary to constitute documentary linguistics as a field of study having the same status as descriptive linguistics. Now, twenty years later, our experience and that of other researchers in South America confirm the intimate feedback relationship between documentation and description. In line with Evans (2008) and

Woodbury (2011), the collaborative preparation of a dictionary and or a grammar can at the same time raise the possibility of eliciting new texts, as shown by the Tapiete (2.1) or Wichí (2.4) experiences. Conversely, the collaborative collection and edition of Ayoreo (2.2) and Niva'ê (2.3) texts constitute a primary source of information to ongoing studies on the phonology and the grammar of those languages. Likewise, community activities carried out during and after the language documentation projects here considered confirm Himmelmann's claim about the creation of available multipurpose corpora which can be used for and beyond linguistic research.

Concerning *participants*, any scientific praxis involving fieldwork with communities cannot and should not avoid reflecting on the impact of the intrusion of field researchers in the lives of community members. This issue becomes relevant in the case of South American indigenous peoples, socioeconomically very vulnerable and historically threatened by non-indigenous society. Faced with this ethical question, the perspective of collaborative fieldwork and the model of empowerment (Cameron et al. 1997) appear as viable alternatives when proposing work agendas agreed on between the researcher and the community, based on common interests.

Furthermore, research on Vilela, the severely endangered Chaco language mentioned above (2.1), highlights the unique role of the last generation of speakers of a language in the documentation, description, and history of their language and people. In addition to the claim by Harrison & Anderson (2008) about the importance of including the speech of semi-speakers and passive speakers in the documentation of endangered languages, we affirm the relevance of applying language-remembering strategies in these situations. As said elsewhere, "the attested processes of linguistic remembering and recovery defy biological metaphors about the vital cycle of languages and their fate" (Golluscio and González 2008: 238).

Finally, the urgent need to document lesser-known still living South American languages, some of them critically endangered, and fill in some remaining gaps in genealogical, typological and areal contact knowledge of the languages makes the establishment of strong collaborative links a central goal. This issue is addressed in the next and final section.

4. Looking ahead: Towards a South American Network of Regional Linguistic and Sociocultural Archives

Different academic institutions in South America involved in linguistic and cultural documentation with indigenous peoples in the region have set up a South American Network of Regional Linguistic and Sociocultural Archives, with the foundational aim of strengthening interaction and exchanging information and documentary resources between archivists, researchers and members of indigenous communities. Some objectives of the agreement include: contributing to the knowledge, preservation, valuation, transmission, and diffusion in national societies of South American languages and non-standard varieties of Spanish and Portuguese, as well as migration languages in contact with them; promoting the development of collaborative language documentation as well as typological and areal research projects; maintaining technological compatibility of files, and, ideally, developing shared criteria for the classification of the contents; implementing a unified code of ethics for researchers, donors, file users, community members and responsible archivists; ensuring long-term preservation of databases and increasing security for stored data; working on creating automatic copies distributed among the archives involved, respecting authorship rights,

ethical principles, regulating access restrictions; and establishing validation processes that can be used for modeling other databases.

This network has its origins in the more than decade-long cooperation between members of South American institutions. The exchange has long been active at informal and personal levels, in particular thanks to researchers participating in DoBeS and ELDP projects, as well as the use of similar technology in individual centers (Seifart et al. 2008). The project currently encompasses CONICET, Universidad Nacional de Formosa and Universidad Nacional de San Juan, Argentina; Instituto de Investigaciones para la Amazonía Peruana; Universidad de Chile; Centro de Estudios Antropológicos of the Universidad Católica de Asunción (CEADUC), Paraguay; Pontificia Universidad Católica del Perú, and Universidad del Azuay, Ecuador. The Museu do Índio and the Universidade Federal do Rio de Janeiro through Bruna Franchetto's active participation, as well as the Museu Paraense Emílio Goeldi, Brazil have shared the Network's objectives from inception (Drude et al. 2009; Golluscio et al. 2013). As of this year, this initiative has been assigned full legal status by an Agreement for Scientific and Technical Collaboration signed by all Network members. The proposal is open to other South American archives or institutions that wish to join this initiative on the understanding that the Network's ethical guidelines are respected.

References

- Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Bem Rampton & Kay Richardson. 1997. Ethics, advocacy and empowerment in researching language. In Nikolas Coupland & Adam Jaworski (eds.), *Sociolinguistics: A reader and course book*, 145–162. Hampshire & London: Macmillan Press Ltd.
- Campbell, Lyle & Verónica Grondona. 2010. Who speaks what to whom? Multilingualism and language choice in Misión La Paz – a unique case. *Language in Society* 39(5). 617–646.
- Campbell, Lyle & Verónica Grondona (eds.). 2012. *The indigenous languages of South America: A comprehensive guide*. Berlin: De Gruyter Mouton.
- CEDIB. (2012). *Indígenas ¿Quién gana, quién pierde?* <http://cedib.org/tag/censo-2012/>.
- Censabella, Marisa. 2009. Capítulo IV. Chaco ampliado. In *Atlas Sociolingüístico de Pueblos Indígenas en América Latina*, 145–169. UNICEF. (<http://www.unicef.org/colombia/centro.htm>).
- Ciccone, Florencia. 2015. *Sustitución lingüística, cambios estructurales y cambios funcionales en tapiete (tupí-guaraní), una lengua en peligro*. Buenos Aires, Argentina: Universidad de Buenos Aires dissertation.
- DGEEC. (2012). *DGEEC: Dirección General de Estadística, Encuestas y Censos*. (<http://www.dgeec.gov.py/microdatos/>).
- Domínguez, Marcelo, Lucía Golluscio & Analía Gutiérrez. 2006. Los vilelas del Chaco: desestructuración cultural, invisibilización y estrategias identitarias. In Lucía Golluscio & Silvia Hirsch (eds.), *Historias Fragmentadas, Identidades y Lenguas: los Pueblos Indígenas del Chaco Argentino*, dossier, *Indiana* 23. 199–226.
- Drude, Sebastian, Bruna Franchetto, Lucía Golluscio, Víctor Miyakawa & Denny Moore. 2009. La red emergente de archivos de lenguas indígenas de la América Latina. Presented at the Meeting “ALICE2, CLARA Europa-A. Latina”, November 18, 2009, Asunción, Paraguay.
- Etacore, Benito & Santiago Durante (eds.). 2016. *Campo Loro gosode oe ojñane udojo – Historias de los pobladores de Campo Loro (edición bilingüe ayoreo-español)*. Buenos Aires: Editorial de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires.
- Evans, Nicholas. 2008. Book Review: Essentials of language documentation. *Language Documentation & Conservation* 2(2). 340–350.
- Fabre, Alain. 1998. *Manual de las lenguas indígenas sudamericanas*, vols 1–2. München & Newcastle: LINCOM EUROPA.
- Fabre, Alain. 2017a [2005]. *Diccionario Etnolingüístico y Guía Bibliográfica de los Pueblos Indígenas Sudamericanos*. (Online edition available at: <http://www.ling.fi/Diccionario%20etnoling.htm>; updated December 2017).
- Fabre, Alain. 2017b. *Gramática de la lengua nivacle (familia mataguayo, Chaco paraguayo)*. LINCOM Studies in Native American Linguistics 78. München: LINCOM.
- Golluscio, Lucía. 2015. Huellas de trayectorias y contactos en el sistema lingüístico: el caso vilela. In Bernard Comrie & Lucía Golluscio (eds.), *Language contact and Documentation / Contacto lingüístico y documentación*, 77–120. Berlin: De Gruyter Mouton.

- Golluscio, Lucía, Bruna Franchetto, Jürg Gasche, Vilacy Galucio & Sebastian Drude. 2013. Weaving a network of archives for South American indigenous languages and cultures. Poster presented at *Language Documentation: Past, Present and Future*. June 5–7, 2013. Hannover, Herrenhausen Castle.
- Golluscio Lucía & Hebe González. 2008. Contact, attrition and shift in two Chaco languages: The cases of Tapiete and Vilela. In K. David Harrison, David S. Rood & Aryenne Dwyer, *Lessons from documented endangered languages* (Typological Studies in Language 78), 195–242. Amsterdam: John Benjamins.
- Golluscio, Lucía & Silvia Hirsch (eds.). 2006. Historias Fragmentadas, Identidades y Lenguas: los Pueblos Indígenas del Chaco Argentino. *Indiana* 23. 97–226.
- Golluscio, Lucía & Alejandra Vidal. 2009–2010. Recorrido sobre las lenguas del Chaco y los aportes a la investigación lingüística. In Lucía Golluscio & Alejandra Vidal (eds.), *Les Langues du Chaco. Structure de la phrase simple et de la phrase complexe*. *Amerindia* 33/34. 3–40.
- González, Hebe. 2005. *A grammar of Tapiete (Tupi-Guarani)*. Pittsburg, PA: University of Pittsburgh dissertation.
- González, Hebe (ed.). 2017. *Ecos del mundo vegetal entre los tapietes de Argentina: diccionario etnobotánico*. LW/D 64. München: LINCOM EUROPA.
- González, Hebe. In press. *Nandipireta ka'a kwawa. Lo que nuestros ancestros sabían del monte. Vocabulario tapiete de plantas*. San Juan: Editorial Fundación de la Universidad Nacional de San Juan.
- Gutiérrez, Analía. 2015. *Segmental and prosodic complexity in Nivaçle: laryngeals, laterals, and metathesis*. Vancouver, BC: University of British Columbia dissertation.
- Hale, Kenneth, Michael Krauss, Lucille Watahomigie, Akira Yamamoto, Colette Craig, Jeanner LaVerne & Nora England. 1992. Endangered Languages. *Language* 68(1). 1–42.
- Harrison, K. David & Gregory D.S. Anderson. 2008. Tofa language change and terminal generation speakers. In K. David Harrison, David S. Rood & Aryenne Dwyer (eds.), *Lessons from documented endangered languages* (Typological Studies in Language 78), 243–270. Amsterdam: John Benjamins.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–95.
- Instituto Nacional de Estadísticas y Censos. INDEC. Encuesta Complementaria de Pueblos Indígenas (ECPI) 2004–2005. Complementaria del Censo Nacional de Población, Hogares y Viviendas 2001. (http://www.indec.gov.ar/principal.asp?id_tema=167).
- Messineo, Cristina. 2008. Fieldwork and speech genres in indigenous communities of the Gran Chaco: Theoretical and methodological issues. *Language Documentation & Conservation* 2(2). 275–295.
- Nercesian, Verónica. 2014a. [2011]. *Wichi lhomtes. Estudio de la gramática y la interacción fonología-morfología-sintaxis-semántica*. München: LINCOM EUROPA.
- Nercesian, Verónica. 2014b. *Manual teórico-práctico de Gramática Wichí*, Vol. 1. Formosa: Editorial de la Universidad Nacional de Formosa.
- Seifart, Frank, Sebastian Drude, Bruna Franchetto, Jürg Gasché, Lucía Golluscio & Elizabeth Manrique. 2008. Language Documentation and Archives in South America. *Language Documentation & Conservation* 2(1). 130–140. <http://hdl.handle.net/10125/1775>.

- Simons, Gary F. & Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the World*, Twenty-first edition. Dallas, Texas: SIL International. Online version. (<http://www.ethnologue.com>) (Accessed 2018-06-23).
- UNICEF. 2010. Los pueblos indígenas en Argentina y el derecho a la educación. Los niños, niñas y adolescentes indígenas de Argentina: diagnóstico socioeducativo basado en la ECPI. http://www.unicef.org/argentina/spanish/4.Libro_ECPI.pdf.
- Unruh, Ernesto & Hannes Kalisch. 2003. Enlhet-Enenlhet. Una familia lingüística chaqueña. *Thule, Rivista italiana di studi americanistici* 14/15. 207–231.
- Vidal, Alejandra. 2001. *Pilagá grammar (Guaykuruan Family, Argentina)*. Eugene, OR: University of Oregon dissertation.
- Vidal, Alejandra (ed.). 2015. *Cuentan los nivaçle Cava chaich'e napi nivaçle*. Formosa: Editorial de la Universidad de Formosa.
- Vidal, Alejandra & José Miranda. 2010. *Diccionario Trilingüe Parlante con ejemplos, notas gramaticales y etnográficas* [Trilingual (Pilagá-Spanish-English) Talking dictionary with examples, grammatical and ethnographic notes]. Formosa: Hans Rausing Endangered Languages Project & Universidad Nacional de Formosa.
- Vidal, Alejandra, Roxana Almeida & José Miranda. 2014a. *Enseñanza de la lengua pilagá. Gramática pedagógica*, vol. 1. Formosa: Editorial de la Universidad Nacional de Formosa.
- Vidal, Alejandra, Roxana Almeida & José Miranda. 2014b. *Enseñanza de la lengua pilagá. Actividades y consignas*, vol. 2. Formosa: Editorial de la Universidad Nacional de Formosa.
- Vidal, Alejandra, Roxana Almeida & José Miranda. 2014c. *Enseñanza de la lengua pilagá. Sugerencias didácticas para el docente*, vol. 3. Formosa: Editorial de la Universidad Nacional de Formosa.
- Vidal, Alejandra & Imme Kuchenbrandt. 2015. Challenges of Linguistic Diversity in Formosa. In Christel Stolz (ed.), *Language Empires in Comparative Perspective*, 89–112. Berlin: Walter De Gruyter.
- Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.
- Zamponi, Raoul & Lucía Golluscio. La lengua vilela en el siglo XVIII. *Simposio: 15/9*. La documentación colonial de las lenguas amerindias: patrones, contextos y aportes conceptuales. Coordinadores: Marisa Malvestitti & Fernando Zúñiga. *56° Congreso Internacional de Americanistas*, Salamanca, 15–20 July 2018.
- Zúñiga, Fernando & Marisa Malvestitti. 2018. Language documentation in the Southern Cone. In Bradley McDonnell, Andrea L. Berez-Kroeker, Gary Holton (eds.), *Reflections on language documentation 20 year afters Himmelmann 1998*, 295–302. (Language Documentation & Conservation Special Publication 15.) Honolulu: University of Hawai'i Press. <http://hdl.handle.net/10125/24830>

Lucía Golluscio
lgollusc@conicet.gov.ar

Alejandra Vidal
vidal.alejandra@conicet.gov.ar

Reflections on fieldwork: A view from Amazonia

Christine Beier

University of California, Berkeley

Patience Epps

University of Texas at Austin

Amazonia is both a place of exceptional linguistic, sociocultural, and ecological diversity and a place where the documentation of this diversity is limited and ever-increasingly urgent. While recent decades have shown considerable progress in this area, our understanding of Amazonian languages is still challenged by a low proportion of researchers relative to its many distinct language contexts. In light of Himmelmann's framing of language documentation as a 'fairly independent field of linguistic inquiry and practice', we discuss key facets of what we consider the single most important unifying question that underlies language documentation work in Amazonia: Just how much description and analysis is necessary for Amazonian language documentation to be coherent, useful, and interpretable by others? We argue that the social and cultural diversity of this vast region calls into question the actual separability of 'documentation' from 'description and analysis' of Amazonian language data; and we advocate for taking Himmelmann's proposals as an invitation to finer-grained, broader-minded thinking about the kinds of research questions, methods, and focused training that best serve linguists working in Amazonian speech communities, rather than as a guide to defining an appropriate scope for fieldwork with an Amazonian language.

1. Introduction Amazonia is a place of exceptional linguistic, sociocultural, and ecological diversity¹—and a place where the documentation of this diversity is both limited and ever-increasingly urgent. At the heart of what Lyon (1974) dubbed the

¹The area encompassing the Amazon and Orinoco river basins is home to some 300 indigenous languages corresponding to over 50 distinct 'genealogical' units, of which the majority are very small families or isolates (see Rodrigues 2000; Epps & Salanova 2013).

“least-known continent”, Amazonia itself was described just twenty years ago as being “still in places a linguistic black box” (Grinevald 1998: 126). While the intervening two decades have seen considerable progress, our understanding of Amazonian languages is still challenged by a low proportion of researchers relative to the many distinct language contexts spread across its roughly 2.9 million square miles. Today, two decades after both Grinevald’s assessment and Himmelmann’s landmark paper on language documentation, a reflection on the state of linguistic fieldwork in Amazonia seems especially fitting.

In this paper, we discuss multiple facets—holistic and conceptual, as well as practical and methodological—of what we consider the single most important unifying question that underlies language documentation work in Amazonia, in light of Himmelmann’s framing of language documentation as a “fairly independent field of linguistic inquiry and practice” (1998: 161) and the still-acute need for high-quality documentation work in the region: Just how much description and analysis is necessary for Amazonian language documentation to be coherent, useful, and interpretable by others?

Speaking from our own experiences working on-the-ground in the Amazonian context, we argue that the social and cultural diversity of this vast region calls into question the actual separability of ‘documentation’ from ‘analysis’ of Amazonian language data. We advocate for taking Himmelmann’s proposals as an invitation to finer-grained, broader-minded thinking about the kinds of research questions, methods, and focused training that best serve a linguist working in an Amazonian speech community, rather than as a guide to defining an appropriate scope for one’s relationship to an Amazonian language.

We begin in §2 by providing some historical context to our discussion. In §3, we highlight key characteristics of the research context in Amazonia; and in §4, we outline key constraints on Amazonian fieldwork. Finally, in §5, we suggest areas to prioritize in future documentation work in Amazonia.

2. Where we’ve been: a brief history of language documentation in Amazonia

Until recent decades, the socio-geographic impenetrability of the Amazonian region limited outside observers to an intrepid, well-funded few, most with non-scientific motivations. Prior to the 1990s, linguistic documentation/description in Amazonia was largely associated with missionary endeavors, from the early Jesuit grammars and catechisms of the 16th and 17th centuries, to the SIL dictionaries and grammars of the 20th century. While valuable, much of this early material is limited in scope and accessibility—for example, dictionaries with dozens of words glossed ‘fish sp.’, grammar sketches in opaque tagmemic framework, and texts limited to Bible translations. Corpora of natural discourse prior to the 1990s are rare and generally limited to a handful of traditional stories. In some cases, more substantial documentation was created by anthropologists, but much of this material lacks linguistically-informed transcription/translation. Vanishingly few materials were made accessible through archiving or as published text collections until quite recently.

The last twenty years have seen major advances in the documentation of Amazonian languages. There has been a significant increase in Latin American scholars working in Amazonia, especially in Brazil, and more foreign scholars have been drawn to the region as well. Increased discipline-wide attention to language documentation has not only stimulated more work; it has also fostered the development of higher standards for documentary collections, including a valorization of rich contextualization and stylistic diversity. Accessibility has also become a priority, and many collections are now widely

available in recently established digital repositories such as AILLA, ELAR, and others (see e.g. Seyfeddinipur et al. forthcoming). Fieldwork in Amazonia has clearly benefited from the international expansion of funding infrastructure, especially the NSF-DEL, ELDP, and DOBES initiatives. These developments have resulted in a relative explosion of high-quality work in Amazonia, including significant text collections, diverse new digital corpora for small and endangered languages, and some excellent descriptive materials—most notably, comprehensive reference grammars grounded in text collections that are openly accessible in digital archives (e.g. Stenzel 2013, Mihas 2015, Zariquiey 2018).

Despite these strides, there is still a tremendous amount of linguistic work to do in Amazonia, and many of the same socio-geographic obstacles remain in place. Many languages still lack basic descriptions, and we have even less information about known types of variation within Amazonian languages—dialects and dialect continua, registers, genderlects, etc.—which demand both documentation and close, context-sensitive analysis, not only to make sense of the variation that occurs within a corpus but also to guide the very process of collection.

At the same time, the contemporary social, political, and economic circumstances of many Amazonian languages make the task increasingly more urgent, as these pressures accompany massive shifts to local *lingua francas*. Moreover, the devastating colonial history of the region—which produced enduring social structures that are deeply devaluing of indigenous languages, knowledge, and lifeways—has left a legacy in which it is often difficult for researchers to establish the trust necessary for respectful and truly collaborative relationships (see also Dobrin & Schwartz 2016).

Our observations here are informed by our many years' collective experience doing linguistic and anthropological fieldwork in Amazonia, as well as training others to work in the region. Our experiences range from work led by a single researcher to team-based projects, involving students and scholars from both outside and inside Latin America, and both closer and looser partnerships with community members.²

3. Building context-sensitive documentation The central proposal of Himmelmann's (1998: 161) discussion is that "documentary linguistics be conceived of as a fairly independent field of linguistic inquiry and practice that is no longer linked exclusively to the descriptive framework." In our view, this proposal is on one hand exactly on target, while on the other hand it requires some important caveats for work with Amazonian languages. Stepping away from the discipline's prior narrow focus on "the descriptive concept of language as a system of units and regularities" (Himmelmann 1998: 164) and toward a broader focus on the whole of a language—within a broader communicative spectrum, which may be multilingual—is essential in the Amazonian context, and our position here is that even the most 'basic' description of a language requires substantial contextualizing work to make it both accurate and comprehensible. But we also submit that the contextualizing work appropriate to Amazonia goes well beyond what many linguists are prepared for. As we elaborate below, the particular features of the Amazonian milieu exhort of us not only a deep awareness of the social and cultural contexts that are home to the language(s), but also a methodological approach that invests in achieving some

²Beier began her long-term relationship with Amazonian peoples and languages in Peru in 1995; Epps in Brazil in 2000. We both are deeply grateful to all of our collaborators and funders over the years, and we take sole responsibility for the views expressed here.

communicative competence,³ makes time for participant-observation within the community, and makes a commitment to ethnography as part of the documentation process.

From our perspective, producing high-quality documentation that is both accurate and interpretable requires of us a coherent *understanding* of the social, cultural, and linguistic contexts in which we are working, both on the intellectual/professional and the ethical/interpersonal levels. This point is relevant in every context, as argued compellingly by Dobrin (2008), who explores a number of foundational ways in which the value systems and priorities of researchers and speech communities can diverge. In Amazonia, at least, we consider it to be a methodological imperative.

Arguably, reaching an appropriate level of understanding may be particularly challenging in Amazonia, where “little-known” (Himmelmänn 1998: 161) languages are generally spoken by ‘little-known’ peoples, whose knowledge systems, value systems, sociopolitical priorities, etc. must be learned, not presupposed. The cultural differences between the local context and a linguist’s background are often very deep, even when the linguist is from the relevant country—and in Amazonia it is very rare for speech community members to lead language documentation projects, especially with a comparable level of training and funding. The process of understanding therefore necessarily involves *analysis* on various levels and with various foci—linguistic, ethnographic, and social. Informed choices about what, when, how, etc. to document depend on this analysis, just as a long-term engagement between a researcher and a community depends on developing mutual understanding to the point that all parties feel comfortable and committed. For example, it is generally expected that a robust documentation of natural discourse will include genres, registers, and styles that are particularly valued by the community; yet sometimes that valuation also corresponds to a heightened sensitivity toward sharing the material with outsiders—whether in light of community norms, negative attitudes on the part of the national society and/or missionaries, or other factors (see e.g. Epps et al. forthcoming).

Thus, for many scholars working in Amazonia, the work of language *documentation* cannot be easily or usefully separated from the work of *description*, just as a focus on *language* cannot be easily or usefully disentangled from an engagement with *culture*. Recognizing that the goals and methods of documentation and description are meaningfully distinct has without a doubt fostered key conceptual innovations in the best practices of our field. However, in light of the shortage of personnel working with any given language or speech community in Amazonia; the likelihood that a linguist’s work may be the first and/or the last work ever done in that setting; and the need for substantial ‘descriptive’ work in order to make a documentation interpretable by others, we have found that the supposed *separability* of these two activities fails to be appropriate in the majority of settings in Amazonia.

4. Practical constraints on documentation It is of course not an accident that Amazonia’s linguistic diversity is severely under-documented. In addition to the cultural challenges that many researchers face, language documentation in the region is confronted by a constellation of practical obstacles.

We turn first to the challenges faced by linguists who come from outside the regional or national context. One set of challenges relates to navigating Amazonian

³In Amazonia, one of the most widespread requirements for building trust is an outsider’s willingness to communicate in the local language.

infrastructures—or the lack thereof—which can be discouragingly difficult, especially for first-timers. The most fundamental steps—getting permissions, gathering resources, getting around over vast distances—are often fraught with complications. In some cases, national policies may actively disfavor or even exclude foreign researchers, often in light of political relations involving their home country.⁴ The day-to-day practical realities of living and working in communities without running water, electricity, or even outhouses can present additional disincentives. In longer-term perspective, even scholars who have carried out successful fieldwork may find it difficult to sustain a research program and collaborative community relationships over time, especially when their home base is far from the region or country where their research takes place.

An additional challenge for many outsiders involves working through a contact language that they do not speak fluently. Most Amazonian languages are spoken in Spanish- and Portuguese-speaking matrix societies; moreover, there are major differences among the varieties of Spanish and Portuguese spoken throughout Amazonia. In Peru, for example, the Spanish of Lima is sufficiently different from the Spanish of rural Loreto that serious attention must be paid to issues of translatability from indigenous languages to the local variety of Spanish to a more internationally-accessible variety of Spanish—and thence to English for most publications. This issue is relevant both to the competent execution of fieldwork and to the nature of its outcomes, including the multiple translations of a documentation necessary for it to be interpretable by multiple audiences.

Linguists who come from within Amazonian regional and national contexts also encounter an array of obstacles to doing language documentation/description. Some of these overlap with those noted above, while many more are structural and financial, varying by country. Crucially, local opportunities for training often do not provide nationally- or regionally-based scholars with the breadth of knowledge, methods, sensitization, tools, funding, professional returns, etc. that they need to do robust documentary/descriptive work.

For linguists or prospective linguists coming from within Amazonian indigenous speech communities, the challenges are in many respects the most daunting. Between local educational realities and national disciplinary priorities, it is extremely difficult and rare for indigenous individuals to pursue advanced education directed toward language documentation. Without such training, it is nearly impossible for them to secure funding, buy equipment, or carry out work according to contemporary standards for best practices. Unfortunately, to this day there are vanishingly few well-trained linguists who are themselves members of Amazonian indigenous communities.

Finally, even those scholars—from any background—who have access to state-of-the-art instruction in linguistics are still unlikely to receive the wide range of training that best serves documentary/descriptive fieldwork in a region like Amazonia. Across the discipline, field methods training is quite limited in scope and duration, and tends to be woefully inadequate on the ‘culture and society’ factors inherent to robust documentation—an issue especially pressing for work with small, under-studied Amazonian societies and speech communities. Yet, because of the political economy of the discipline, there is rarely an easy way to offer significantly greater depth and breadth of training. For formally-trained linguists who choose to branch out into language

⁴For example, Venezuelan languages are among the least-documented, due in part to Venezuelan national policies regarding researchers from a range of countries on the one hand, and relative lack of support or training in documentation for local scholars on the other.

documentation/description later in their career, the availability of thorough methods training is likely to be even less.

5. Priorities for the future In light of the challenges addressed above, we outline here what we see as important priorities for the future of language documentation, with emphasis on the Amazonian context.

Disciplinary priorities. In our view, documentation/description activities are still sufficiently undervalued in the discipline of linguistics as a whole that even linguists (especially graduate students) who are interested in working in Amazonia sometimes decide not to take a “professional risk” with a long-term commitment to this type of scholarship. Even in linguistic departments like UC Berkeley’s or UT Austin’s, where commitment to description and documentation is both historically foundational and currently vibrant, graduate students who are primarily interested in these areas encounter structural and even attitudinal obstacles in the course of their training. Post-degree employability is a major concern for students and their advisors alike; normative time expectations and disciplinary conventions regarding what ‘counts’ as a dissertation topic can strangle documentation and even heavily descriptive projects; and clashes between sub-disciplinary values and priorities can be more corrosive than many of us realize. Moreover, since there are presently no avenues for long-term stable employment as a Language Documentarian, such work is either secondary to teaching or is short-term and project-based, as in the case of post-doctoral positions. Yet the progress that an individual (or even a team) can make documenting a small Amazonian language in a non-urban setting, when limited to academic summers, is discouragingly slow.

Training and expectations for fieldwork. Given the constellation of factors unique to the region, there is a clear need for more field schools in Amazonia; and for more team-based research projects involving *in situ* apprenticeship components. Many challenges that we have discussed here could be effectively addressed in the context of collaborative, ‘inter-generational’ training opportunities in Amazonia, especially in partnership with local universities in cities like Iquitos or Leticia. Building in more time, academic credit, institutionally-supported programs, durable funding opportunities, and higher standards for focused *in situ* training could transform the quality of both the experience of Amazonian fieldwork and its tangible outputs. Less ambitiously, more and better training within existing field methods courses in areas including cross-cultural sensitivity, participant-observation, ethnography, and archiving would better equip budding Amazonianists with the range of skills they need for creating appropriate, accurate, and interpretable language documentations.

At the same time, because of the conditions specific to documentation work in Amazonia, it seems crucial that basic disciplinary expectations become more realistic regarding how much time, training, and resources are necessary for good documentation work. This is relevant in multiple domains, but especially in gauging how much output a specific field project or fieldwork period is designed to accomplish; how much training researchers get as cross-cultural, multi-lingual fieldworkers; how much time and breadth advisors and students carve out for graduate-level research in Amazonia; how long it ‘should’ take to write a good dissertation about an Amazonian language; and what kinds of work ‘count’ toward tenure.

Engaging with the documentation of speech practices and with ethnography.

Because speech practices may be variable and even multi-lingual in a single small setting, as is often encountered in Amazonia, the one-language focus that is typically assumed as a standard for documentation is, in some contexts, artificial and not ethnographically appropriate. Similarly, any documentation of an Amazonian language that could be defined as ‘comprehensive’ will require significant culturally- and socially-contextualizing components. Many linguists now understand the importance of incorporating ethnographic work into their research; similarly, many anthropologists now recognize the methodological flaws of working exclusively through a contact language. At the same time, however, including variation and ‘thick description’ in documentation introduces significant additional challenges, notably, the need to balance realistic temporal and material constraints on a single project while engaging with the richness that is discovered in the context. Again, these issues exhort us to recalibrate our expectations.

Ethics and collaboration. In the Americas as elsewhere, there is a growing sensibility that linguistics must not be an ‘extractive’ enterprise. One outcome of this new awareness has been the emergence of more genuinely collaborative efforts between linguists and speech community members. In our view, this is a hugely positive development in the relationship between ‘linguistics’ and the rest of the world. At the same time, on a practical level, this means that linguists now share control, timelines, resources, outcomes, etc. with their collaborators in ways that often clash with disciplinary expectations and structures. Many collaboration-oriented fieldworkers struggle to integrate the inward-facing facets of academic linguistics with their commitments to responsible outward-facing work, plus the ample time, commitment, dedicated resources, and energy required to do their work well.

Preservation and sharing of documentation. Despite gains in the last twenty years, there is still a great need for educating people about the importance of archiving their materials, as well as exactly how to go about it. This need includes greater attention to regional or national contexts, where people are often hesitant to let materials go into an archive based outside the country, and yet there is no viable local option. The legacy of Amazonian languages depends on more effective dissemination of, and recognition of, the products of good documentation work, so that the data can be better used by non-fieldworking linguists. It also depends on the accessibility and usefulness of these materials to communities, where they may contribute to maintenance and revitalization efforts, and represent a resource for future generations. These considerations underscore the need to build greater recognition for archived materials within the field, as well as for community-directed outputs such as pedagogical materials.

In conclusion, language documentation in Amazonia comes with particular challenges, but also with particular rewards. The region’s diversity of languages offers a seemingly endless array of surprises for linguistic theory and typology; cultural differences provide us with new opportunities to discover how human beings engage with their social and ecological worlds; and the documentary enterprise supports speakers in maintaining their heritage and strengthening their position *vis-à-vis* national and global societies. The Amazonian context underscores the need to develop new and more holistic approaches—


both in our thinking and in our methods—that span documentation and description, and to engage them from both linguistic and ethnographic perspectives.

References

- Dobrin, Lise M. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment in Melanesia. *Language* 84(2): 300–324.
- Dobrin, Lise M. & Saul Schwartz. 2016. Collaboration or participant observation? Rethinking models of 'Linguistic Social Work'. *Language Documentation & Conservation* 10, 253–277.
- Epps, Patience & Andrés Pablo Salanova. 2013. The languages of Amazonia. *Tipiti: Journal of the Society for the Anthropology of Lowland South America* 11(1): 1–27.
- Epps, Patience L., Anthony K. Webster & Anthony C. Woodbury, forthcoming. Documenting speech play and verbal art: A tutorial. *Language Documentation & Conservation*.
- Grinevald, Collette. 1998. Language endangerment in South America: A programmatic approach. In Lenore Grenoble & Lindsay Whaley (eds.), *Endangered languages: Language loss and community response*, 124–260. Cambridge: Cambridge University Press.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1): 161–195.
- Lyon, Patricia J. (ed.). 1974. *Native South Americans: Ethnology of the Least Known Continent*. Prospect Heights, Illinois: Waveland Press, Inc.
- Mihas, Elena. 2015. *A grammar of Alto Perené (Arawak)*. Berlin: de Gruyter Mouton.
- Rodrigues, Aryon D. 2000. Panorama das línguas indígenas da Amazônia [Overview of the indigenous languages of Amazonia]. In F. Queixalós & O. Renault-Lescure (eds.), *As línguas amazônicas hoje* [The Amazonian languages today], 15–28. São Paulo: Instituto Socioambiental, Museu Paraense Emílio Goeldi.
- Seyfeddinipur, Mandana, Felix Ameka, Jonathan Blumtritt, Lissant Bolton, Irmgarda Kasinskaitė-Buddeberg, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience Epps, Vera Ferreira, Ana Vilacy Galucio, Birgit Hellwig, Oliver Hinte, Gary Holton, Maja Kominko, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Hubertus Neuhausen, Sebastian Nordhoff, Brigitte Pakendorf, Felix Rau, Keren Rice, Michael Rießler, Nick Thieberger, Paul Trilsbeek, Hein Van der Voort, Kilu von Prince & Anthony Woodbury. forthcoming. Public Access to Research Data in Language Documentation: Challenges and possible strategies for dealing with them. *Language Documentation & Conservation*.
- Stenzel, Kristine. 2013. *A reference grammar of Kotiria (Wanano)*. Lincoln: University of Nebraska Press.
- Zariquiey Biondi, Roberto. 2018. *A grammar of Kakataibo*. Berlin: de Gruyter Mouton.


Christine Beier

beiercm@berkeley.edu

 orcid.org/0000-0002-2741-2676

Patience Epps

pattiepps@austin.utexas.edu

 orcid.org/0000-0002-7429-7885

Reflections on linguistic fieldwork in Mexico and Central America

Gabriela Pérez Báez
University of Oregon
Smithsonian Institution

In this chapter, I endeavor to contribute towards a collective effort to reflect on the evolution and state-of-the-art of language documentation. I reflect on Himmelman (1998) from the perspective of language endangerment and revitalization in Mexico and Central America today. I identify a number of topics that are critical to the practice of language documentation in the region and that in my view were only marginally mentioned in Himmelmann's seminal paper. These topics revolve around the participation, consent, interests and needs of speakers of the very languages that are documented. Notably, I argue that (i) language documentation is critical for language revitalization, (ii) I echo current calls in the community-based research literature for ensuring that language documentation is collaborative, (iii) that to this end, training opportunities for language community members need to increase and (iv) that a concerted effort is needed to develop appropriate ways to ensure informed consent in language documentation.

1. Introduction¹ In this chapter, I endeavor to contribute towards the collective effort in this volume to reflect on the evolution and state-of-the-art of language documentation. I hope to bring to the forefront topics that are especially relevant to Mexico and Central America today. Pérez Báez, Rogers and Labrada (2016) analyze the many factors that impact language documentation as practiced in Latin America and argue for the need to develop practices and principles that are in line with the particulars of Latin American contexts. It is with the hope of contributing towards fulfilling this need that I write this chapter.

¹I am grateful for comments and suggestions from Hilaria Cruz Cruz, Alí García Segura, Carolyn O'Meara, Carlos Sánchez Avendaño, Mandana Seyfeddinipur, and Susan Smythe-Kung. Any errors or omissions are of course my sole responsibility.

2. Usability and language documentation in the context of language endangerment An argument that Himmelmann (1998) makes in support of language documentation as an independent, stand-alone endeavor is that “Collections of primary data have at least the potential of being of use to a larger group of interested parties” (p. 163). This argument raises questions about the extent to which primary data collected without a particular hypothesis as a goal can inform an analytical pursuit. I do not wish to expound on this particular point, however. Rather, I focus on the potential that language documentation does have for one particular group—that of the speakers of the documented language. Himmelmann does state that among potential beneficiaries of language documentation is “...the speech community itself, which might be interested in a record of its linguistic practices and traditions” (p. 163). In hindsight, and in the context of the severity of the global language endangerment crisis and the actions of the linguistics community in response to it, the marginal mention in Himmelmann’s article of the speech community as a beneficiary of language documentation demands attention.

Data from the Endangered Languages Catalog (ELCat) include 164 languages in Mexico and Central America in some stage of language endangerment (Table 1). The Global Survey of Revitalization Efforts (henceforth the Survey) (Pérez Báez, Vogel and Okura 2018, Pérez Báez, Vogel & Patolo, in press) documented 245 efforts around the world: 32 were in Mexico and five in Central America. Table 2 shows that 18 efforts were for languages that have now lost their child speakers (categories 1, and 4 to 8) while 14 have less and less children (categories 5 and 6). Only 8 languages still have child speakers (categories 6 and 7).

Endangerment status	Speaker number trends	Number
Dormant	There are certain languages about which one source says the language in question is “extinct,” “probably extinct,” “possibly extinct,” or has “no known speakers,” where another equally credible source reports it as still having speakers. In this Catalogue (ELCat), languages of this sort as well as languages whose last fluent speaker is reported to have died in recent times, even when sources do not disagree are listed as “dormant.”	6
Awakening	Languages which have lost their last native speakers but which have on-going revitalization efforts	2
Critically Endangered	A small percentage of the community speaks the language, and speaker numbers are decreasing very rapidly.	9
Severely Endangered	Less than half of the community speaks the language, and speaker numbers are decreasing at an accelerated pace.	6
Endangered	Only about half of community members speak the language. Speaker numbers are decreasing steadily, but not at an accelerated pace.	17
Threatened	A majority of community members speak the language. Speaker numbers are gradually decreasing.	58
Vulnerable	Most members of the community or ethnic group speak the language. Speaker numbers may be decreasing, but very slowly.	49
At risk ²		13
No LEI		4
TOTAL		164

Table 1: Language endangerment in Mexico and Central America. (Source: <http://www.endangeredlanguages.com> last accessed on April 14, 2018.)

²An *at risk* language is one with an LEI of 0 but for which the confidence is lower than 100%, meaning that not all factors that determine the LEI are known (Holton, p.c., April 14, 2018).

Status	Number
1 There are no first-language speakers.	2
2 There are a few elderly speakers.	2
3 Many of the grandparent generation speak the language, but the younger people generally do not.	8
4 Some adults in the community are speakers, but the language is not spoken by children.	6
5 Most adults in the community are speakers, but children generally are not.	10
6 Most adults and some children are speakers.	4
7 All members of the community, including children, speak the language, but we want to make sure this doesn't change	4
8 There is a new population of speakers or people are beginning to learn the language after a period of time in which no one spoke the language.	0
Regional subtotal for Mexico and Central America	36

Table 2: Intergenerational transmission index for Central America and Mexico in the Global Survey of Revitalization Efforts

In reference to language endangerment, Himmelmann states: “My concern is the application of this framework for recording little-known or previously unrecorded languages. Most of these languages are endangered...” (p. 176). This focuses on addressing the likelihood that many languages of the world would cease to be available for future study. The data I have presented above makes the severity of the endangerment situation quantifiable and supports Himmelmann’s call. I wish to add, however, that language documentation is of high importance to those engaged in revitalizing their languages.

In the Survey, respondents had the opportunity to articulate, in open text fields, up to five revitalization objectives. Nine categories were identified and responses were coded into them (see Pérez Báez, Okura and Vogel 2018 and Pérez Báez, Vogel & Patolo in press for coding methods). The top objectives category documented at a global level was language teaching with 23.7% of responses. Language documentation and analysis came in fourth place with 13.4% of responses. Regionally, in México and Central America language teaching was also the top category with 20% of responses but language documentation and analysis came in second place with 18% of the objectives articulated. One respondent states as an objective *Rescatar y documentar la lengua a través de entrevistas con hablantes [e] investigaciones históricas y lingüísticas* (‘To revive and document the language by way of interviews with speakers and historical and linguistic research’). Another focuses on *Conformación de materiales didácticos para la enseñanza del hñáñho* (‘creation of pedagogical materials to teach Hñáñho’). Whether documentation is explicitly stated as an objective or we infer that documentation is needed for creating language teaching materials, these numbers provide us with empirical data to unequivocally argue that language documentation is critical for language revitalization.

3. Participation of and consultation with native speakers Related to the relevance of language documentation for revitalization has been the role of members of the language community in the documentation process itself. A considerable shift in the last 20 years is the increasing advocacy for community-based research in linguistics. Himmelmann considers “close cooperation with members of the speech community” (p. 171) as necessary for high-quality documentation and that “Ideally, the person in charge of the compilation speaks the language fluently and knows the cultural and linguistic practices in the speech community very well” (p. 171). It is then from within the language community that the better suited compilers of language documentation may be found. Bribri speaker, researcher and cultural expert Alí García Segura has published on ecological knowledge and takes the time to provide non-Bribri readers with explanations that might bring them closer to Bribri thinking (see García Segura 2016). Cruz (2017) provides an outstandingly detailed and insightful analysis of *La42 qin4 kchin4* (“Prayers for the Community”), “part of a ritual carried out regularly by elders and traditional San Juan Quiahije authorities in their official capacity as community representatives” (509). As a member of the community and native speaker of the Quiahije Chatino language, Cruz is not only strongly positioned to interact with the elders and authorities for the purpose of obtaining permission to document the prayers but has the cultural acumen to carry out a valid and valuable verbal art analysis (see Chapter 37, this volume). When the life of Emiliano Cruz Santiago was cut short, not only was it a painful loss to those who love him, but also his native Southern Sierra Zapotec language lost a dedicated and meticulous documenter of the cultural knowledge embedded in it (see Cruz Santiago 2010). The level of cultural acuity that documenters such as García Segura, Cruz and Cruz Santiago possess is likely well beyond the reach of any researcher external to the language communities.

However, community collaboration or the active participation in language documentation by native speakers cannot solely be motivated by a drive for quality assurance.

Ethical considerations in linguistic research and advocacy for community-based research (CBR) in 1998 were already articulating the critical nature of participation of, and consultation and collaboration with members of a language community. Himmelmann cites seminal works on these topics and asks “how the communities can be actively involved in the design of a concrete documentation project” (p. 188). It is noteworthy that literature on CBR has been influenced by research in Central America that predates Himmelmann (1998). Nora England, based on experiences with speakers of Mayan languages in Guatemala, delivered a poignant statement about the obligations of linguists with regards to language community members in Hale et. al (1992). Other examples come from experiences working with the Rama (Grinevald and Pivot 2014 *inter alia*) and the Sumu-Mayangna communities (Benedicto et. al 2007 *inter alia*) both in Nicaragua. Fast forward to the year 2018 and the literature on CBR has become copious with works such as the recent volume *Perspectives on Language and Linguistics: Community-Based Research* (Bischoff and Jany 2018), with three chapters based on Mexico and Central America.

Alongside collaboration has come advocacy for the training in documentation for native speakers of languages that are endangered and/or not well documented, both within and outside degree-granting programs. At the doctoral level, Mexico’s Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS) focuses on “*La formación de especialistas hablantes de una lengua indígena que puede ser su primera o segunda lengua, siempre y cuando acrediten conocimiento de la lengua indoamericana que vayan a trabajar*” (“specialized training for speakers of an indigenous language, be it their first or second language, provided they are able to show knowledge of the Indo-American language they intend to work with”).³ Outside of Mexico, programs such as that of the University of Texas at Austin have trained generations of speakers of indigenous languages from the Americas who now comprise a cadre of excellent documenters, researchers and faculty members practicing in top universities. Academic institutions are increasingly active partners of language communities for documentation. One example is the Escuela de Filología, Lingüística y Literatura at the Universidad de Costa Rica, which combines higher-education training and research into a social action model based on close collaboration with the language communities of the country. This approach has led to teacher training programs and a copious production of teaching materials and dictionaries.

Outside degree-granting institutions, a recent example is the language documentation workshop offered in Mexico in December 2017 by the Endangered Languages Documentation Program (ELDP) in collaboration with CIESAS and the Biblioteca de Investigación Juan de Córdova. The training was offered primarily to native speakers of Mesoamerican languages. There is high-demand for training of this type. For instance, community members may not wish or be able to leave their communities for a number of years in order to acquire language documentation skills through a degree-granting program. This is especially the case for those who have government responsibilities in their communities. Similarly, teachers with a degree in hand and an ongoing career might only be able to meet their needs for training by attending workshops offered outside working hours or outside the academic year. Pérez Báez (2016) describes language revitalization activities

³From the call for admissions applications to the 2018-2022 program at <https://docencia.ciesas.edu.mx/wp-content/uploads/2017/12/COMP-Cartel-Linguistica-1-1.pdf>, last accessed on April 19, 2018. The translation is my own.

by school teachers in the Zapotec community of San Lucas Quiavini. A couple of months prior to the writing of this paper, some of the teachers reached out to inquire about opportunities to obtain equipment, funding and training for language documentation as part of the school's activities. The opportunities within country, it turns out, are very limited. Outside of Mexico, opportunities such as those offered by the Endangered Language Fund and the Foundation for Endangered Languages for funding, and the longstanding Institute on Collaborative Language Research (CoLang) for training, are of the scope to serve some of these needs. However, the dependency on English as the working language of these programs stands as a solid barrier for many in Mexico and Central America.

4. Ethics, consent and access Himmelmann raises the issue of community rights over the documentation and dissemination of language data. This topic has developed a stronghold in the practice of language documentation, as it should have. However, guidelines for ethics in language documentation in Mexico, for instance, are not well established, and neither is the process of institutional review of a project or informed consent. With national and international researchers carrying out documentation in Mexico, there are inconsistencies in the process by which informed consent is obtained. Further, with institutional review boards outside of Mexico dictating ethical protocols for research in Mexico, questions arise about the influence and appropriateness of such principles.

A related issue is that of ensuring that language community members will have access to the products of language documentation projects. O'Meara and González Guadarrama (2016) expound on the complications they found in creating community access to language documentation in Seri and Nahuatl communities in Mexico. In recent years, the Red de Archivos de Lenguas México (RALMEX) emerged as a network of institutions involved in language documentation. CIESAS, in collaboration with the Dokumentation Bedrohter Sprachen (DoBeS) program at the Max Planck Institute for Psycholinguistics and LinguaPax, created the Acervo Digital de Lenguas Indígenas Víctor Franco Pelletier, which provides access to materials in a dozen language groups. I venture to say, however, that the bulk of the documentation produced in keeping with contemporary best practices in language documentation is archived in two primary repositories: The Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin and the Endangered Languages Archive (ELAR) at SOAS University of London. These archives have an international scope and are faced with the challenge of making their documentation accessible in *lingua francas* other than English. AILLA offers an interface in Spanish in addition to English. ELAR's interface is in English but the idea of developing portals for communities in the language of a particular documentation deposit in addition to the major *lingua franca* has been under consideration.

5. Conclusions In this contribution I have reflected on the role and relevance of language documentation in the linguistic landscapes of Mexico and Central America from the vantage point of 20 years after the publication of Himmelmann (1998). I have identified a number of topics that are critical to the practice of language documentation in the region and that in my view were only marginally mentioned in Himmelmann's seminal paper. These topics revolve around the participation, consent, interests and needs of speakers of the very languages that are documented. I hope to have argued convincingly that


these topics demand that the linguistics community, and anyone involved in language documentation, keep them in the forefront.

References

- Benedicto, Elena, Demetrio Antolín, Modesta Dolores, M. Cristina Feliciano, Gloria Fendly, Tomasa Gómez, Baudillo Miguel & Elizabeth Salomón. 2007. A model of participatory action research: The Mayangna linguists' team of Nicaragua. In Maya Khemlani David, Nicholas Ostler and Caesar Dealwis (eds.), *Proceedings of the XI FEL Conference on 'Working together for endangered languages - research challenges and social impacts'*, 29–35. Kuala Lumpur: SKET, University of Malaya and Foundation for Endangered Languages.
- Bischoff, Shannon & Carmen Jany (eds.). 2018. *Perspectives on Language and Linguistics: Community-Based Research*. Amsterdam: De Gruyter Mouton.
- Cruz, Hilaria. 2017. Prayers for the Community: Parallelism and Performance in San Juan Quiahije Eastern Chatino. *Oral Tradition* 31(2). 509–534.
- Cruz Santiago, Emiliano. 2010. *Jwá'n ngwan-keéh reéh xa'gox* (Creencias de nuestros antepasados). Colección "Diálogos, Pueblos Originarios de Oaxaca". Oaxaca: Culturas Populares.
- García Segura, Alí. 2016. *Ditsö` rukuö`* (Identity of the seeds: Learning from nature). Gland: International Union for Conservation of Nature.
- Grinevald, Colette and Bénédicte Pivot. 2014. On the revitalisation of a 'treasure language': the Rama language project of Nicaragua. In M. Jones & S. Ogilvie (eds.), *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, 181–197. Cambridge: Cambridge University Press.
- Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne & Nora C. England. *Language* 68(1). 1–42
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195.
- Lee, Nala H. and John Van Way. 2016. Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI). *Language in Society* 45.2. 271–292.
- O'Meara, Carolyn and Octavio Alonso González Guadarrama. 2016. Accessibility to results and primary data of research on indigenous languages. In Gabriela Pérez Báez, Chris Rogers & José Emilio Rosés Labrada (eds.), *Latin American contexts for language documentation and revitalization*, 30–59. Amsterdam: De Gruyter Mouton.
- Pérez Báez, Gabriela. 2016. Addressing the gap between community beliefs and priorities and researchers' language maintenance interests. In Gabriela Pérez Báez, Chris Rogers & José Emilio Rosés Labrada (eds.), *Latin American contexts for language documentation and revitalization*, 165–194. Amsterdam: De Gruyter Mouton.
- Pérez Báez, Gabriela, Rachel Vogel & Eve Okura. 2018. Comparative analysis in language revitalization practices: Addressing the challenge. In Kenneth L. Rehg & Lyle Campbell (eds.), *Oxford Handbook of Endangered Languages*, 466–489. Oxford: Oxford University Press.
- Pérez Báez, Gabriela, Rachel Vogel & John Uia Patolo. In press. Global Survey of Revitalization Efforts: A mixed methods approach to understanding language revitalization practices. *Language Documentation & Conservation*.

Pérez Báez, Gabriela, Chris Rogers and Jorge Emilio Rosés Labrada. 2016. Introduction. In Gabriela Pérez Báez, Chris Rogers & José Emilio Rosés Labrada (eds.). *Latin American contexts for language documentation and revitalization*, 1–28. Amsterdam: De Gruyter Mouton.

Gabriela Pérez Báez
gperezb4@uoregon.edu

 orcid.org/0000-0003-2870-2306

Reflections on language documentation in North America

Daisy Rosenblum

University of British Columbia

Andrea L. Berez-Kroeker

University of Hawai'i at Mānoa

In this paper we reflect on the state of language documentation in North America, especially Canada and Alaska. Using our own early experiences with the archival record on languages of North America as a launching point, we discuss changes that have come to this field over the past twenty years. These include especially the increasing recognition of long traditions of community-based language research within North America, and of members of language communities as primary stakeholders in efforts to preserve and properly share records of linguistic knowledge.

In this paper we reflect on the state of language documentation in North America, especially Canada and Alaska. Using our own early experiences with the archival record on languages of North America as a launching point, we discuss changes that have come to this field over the past twenty years. These include especially the increasing recognition of long traditions of community-based language research within North America, and of members of language communities as primary stakeholders in efforts to preserve and properly share records of linguistic knowledge.

For each of us, our own first encounters with language documentation led us to understand, appreciate, and ultimately strive to practice community-engaged and community-directed research (eg., Czaykowska-Higgins 2009). Our personal trajectories as researchers align with and reflect a paradigm shift around language research in the United States and Canada, also echoing changes in other parts of the world with shared colonial histories persisting in present realities. In this chapter, we describe what we see emerging as standards of practice in North America. We also tell our origin stories as researchers working with language and community, and in so doing, we adopt methodology we have learned from our Indigenous research partners: to introduce

ourselves, to explain where we are from, and why we are doing this work. We begin at the end: with the products of language documentation projects stored in archives.

Daisy: *Twenty years ago, around the time Himmelmann 1998 was published, I visited the Rare Book and Manuscript Library at the Butler Library of Columbia University. I was an undergraduate student in Sally McLendon’s Hunter College class on North American Languages and Cultures, and my assignment was to find an original manuscript in an Indigenous American language to write about. An archivist placed a box on the table in front of me and I pulled out a stack of manuscripts pencilled in George Hunt’s careful hand. I read through the texts and stopped at one titled ‘The Brothers’, a Comox story written in Kwak’wala. In the interlinear translation, I recognized a Wakashan version of a Salishan story written about by Dell Hymes in “In Vain I tried to Tell You”, with telltale motifs of spousal betrayal, transvestite deception, brotherly revenge, and a younger sibling who tries but fails to warn her family of danger in the home (Hymes 1981: 317). I didn’t yet know what language documentation was, nor that the papers I was reading belonged to a ‘Boasian trilogy.’*

In defining documentary linguistics as a separate pursuit from descriptive linguistics, Himmelmann proposed that language documentation be conceived of as a “radically expanded text collection...suitable for a range of purposes.” Himmelmann does not mention Boas, but the prototypical documentations in North America are Boasian trilogies of a grammar, dictionary and set of texts, created by Boas himself or one of his many students.¹ Himmelmann encourages linguists to value documentation and see, as noted by Rice 2011, that “it is impossible to imagine current linguistic theory, be it formally or functionally oriented, without the existence of the quality descriptions found in the Boasian trilogy” (Rice 2011: 192). Such Boasian trilogies originated in a moment of salvage ethnography, born of the presumptive nostalgia assigned to Native communities imagined to be in the process of disappearing. And yet, twenty years after Himmelmann, we can see that the greatest value of good documentation is to today’s descendants of the speakers themselves.

Daisy: *I wrote a paper about Kwak’wala discourse markers in ‘The Brothers’, and the erasure of these discourse markers in published versions of the story.² Ten years later, that paper was part of my graduate application to the UC Santa Barbara. The following summer I found myself working with two speakers of Kwak’wala, Beverly Lagis and Daisy Sewid-Smith, and several community members engaged in language documentation and reclamation, during the inaugural InField in a class coordinated and led by Patricia A. Shaw. A collaboration with Mikael Willie from Kingcome Inlet, a language and culture teacher who participated in the course, brought me to the Tsulquate Reserve of the Gwa’sala-’Nakwaxda’xw Nations, where I continue to work today in partnership with the Elementary School on the Reserve.*

The research partnership between George Hunt and Franz Boas produced copious documentation of Kwak’wala language and Kwakwaka’wakw culture; some of these records remain at the Columbia University where Boas founded the Department of Anthropology. Another large set of records is archived at the American Philosophical Society (APS) in Philadelphia. Brian Carpenter, the curator of Native American Materials at the APS, described recent news of the collection in March 2018:

¹See https://en.wikipedia.org/wiki/Franz_Boas#Students_and_influence for a list of Boas’ students and their influence.

²The image in Figure 1 is not of *The Brothers* but shows a manuscript from the same collection; the Boas fonds at the Columbia RBML are still not digitized, but the Library provided an image of another manuscript.

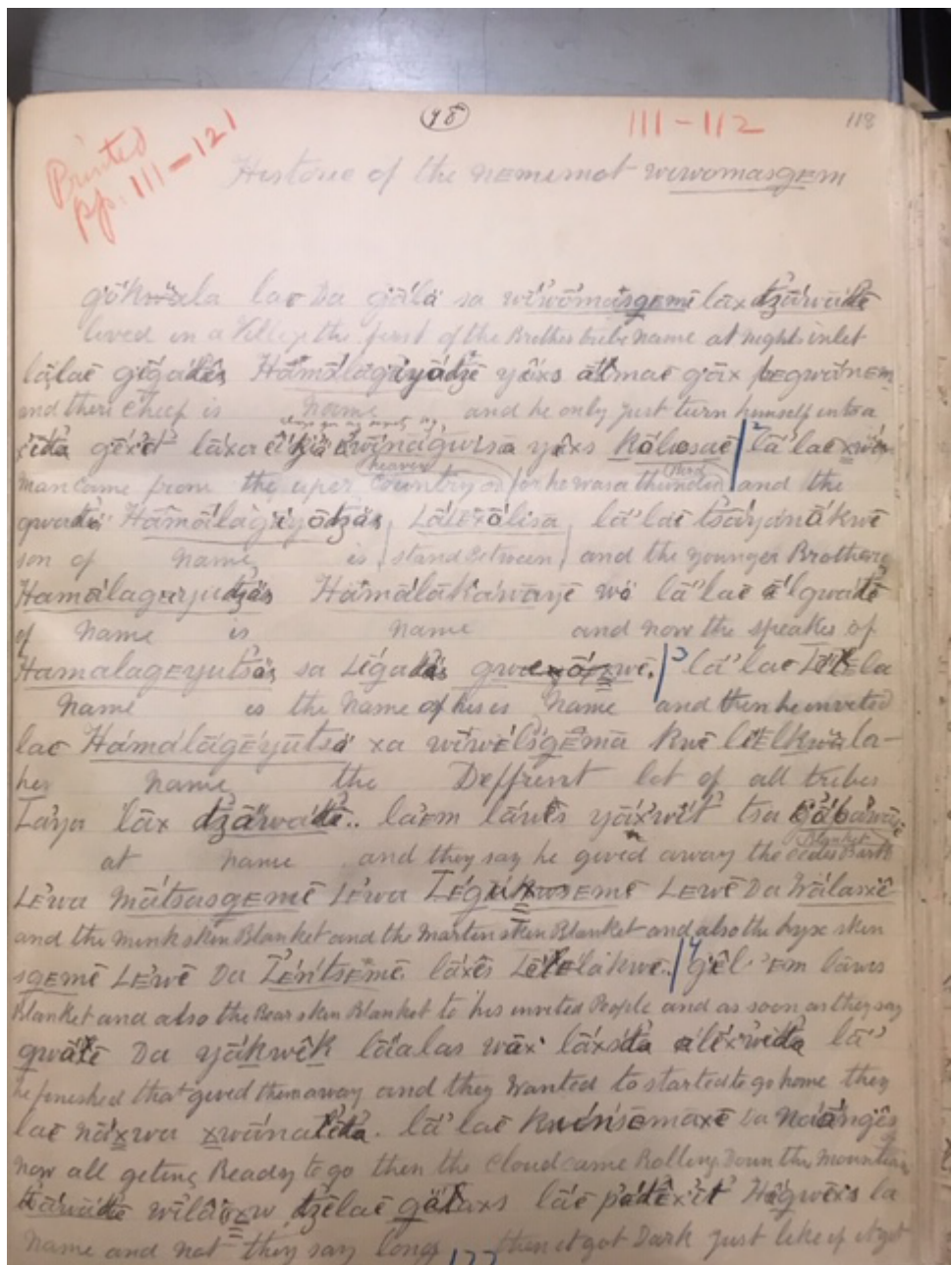


Figure 1: A manuscript page written in George Hunt's hand, with added annotations by Franz Boas. Photo credit: Tara Craig.



Figure 2: Gwi'molas potlatch.

(i)n 2015 and 2016, the APS was honored by an invitation to attend two Kwakwaka'wakw potlatches held in Alert Bay, British Columbia. At these traditional gatherings, the APS gave away books [...] containing unpublished manuscripts from the APS Library written by George Hunt in the Kwak'wala language and English. These books [...] were distributed to the assembled hereditary chiefs, matriarchs, singers, and other community members. It is precisely these people who are the primary constituency, core experts, and research public for these manuscripts. (Carpenter 2018; italics added)

Carpenter continues:

This distribution of just some of Hunt's materials served as the largest increase in access to these materials for the Kwakwaka'wakw community in history. Most importantly, this approach came about entirely through the guidance of Kwakwaka'wakw people. ... [We are] working with members of several Kwakwaka'wakw communities on making the contents of Hunt's ... manuscripts easier to navigate, and also seeking guidance to help ensure that the materials are represented and utilized in ways that are respectful of Kwakwaka'wakw protocols concerning [...] different kinds of knowledge and information. (Carpenter 2018)

For Carpenter, the Kwakwaka'wakw community members attending the potlatch and receiving copies of George Hunt's work are the '*primary* constituency, core experts

and research public' for the materials contained in a language documentation. This is not a necessarily a new phenomenon: there has always been strong interest among community members in the work carried out by linguists and anthropologists about their language and culture. While Himmelmann acknowledged the likely interest of community members in documentation, his definition of language documentation was oriented toward an audience of university-based researchers interested in questions about Language and linguistic structure. However, as indicated by Carpenter, it turns out that within North America the most significant and core audience which one may anticipate will access documentations of a language is *not* other linguists interested in typological or theoretical questions (who remain a relatively small group), but the members of a given speech community interested in researching, learning and teaching their heritage languages, many of whom may be accessing recordings of their ancestors.³ These numbers are growing, as exemplified in the success and growth of the 'Breath of Life' model of archival research, initiated at University of California Berkeley by Advocates for Indigenous California Language Survival over twenty years ago and since replicated in Washington D.C., Oklahoma and British Columbia.⁴ As a result, archives play a significant role in linking community members to linguistic and cultural knowledge. Existing repositories are thus reconceiving their relationship to content, while new repositories are being developed under community control.

Andrea: *My entree into the world of language documentation in North America began five years after Himmelmann 1998 was published. In 2003 I was an MA student in linguistics and a student employee of the LINGUIST List, and I was selected, along with my coworker Sadie Williams, to participate in an NSF-funded project known as the Dena'ina Archiving, Training and Access project (DATA, Holton et al. 2006). Under the direction of the project PI, Gary Holton, our task was to travel to Alaska to assist in developing an online database of the Dena'ina language holdings in the Alaska Native Language Center. The ANLC was at the time just beginning its foray into digitizing their massive paper-and-tape collection, guided by the best practice recommendations of another LINGUIST List project known as E-MELD (Electronic Metastructure for Endangered Languages Data; Boynton et al. 2006).*

The three semesters and two summers I spent working on the DATA project were undeniably formative. I witnessed first-hand the tremendous impact that access to records of one's own linguistic inheritance had on Dena'ina Elders and young language learners alike. Later, during my doctoral studies, Kari Shaginoff of Nay'dini'aa Na' Kayax (Chickaloon Village) invited me to the land of the neighboring language Ahtna, where Karen Linnell, Liana Charley and Taña Finnesand of the Ahtna Heritage Foundation brought me onto their team to build C'ek'aedi Hwnax, the Ahtna language digital archive that is fully administered by the nonprofit wing of Ahtna, Incorporated⁵ (Berez et al. 2012).

These early experiences with these various models of archiving reflect a more general trend in archiving of North American languages, in which a shift in the locus of archiving practice is slowly becoming Indigenous-community centered. The ANLC was for many years an excellent example of the kind of institutionally-based brick-and-mortar repositories that dominated the archiving landscape in the 19th and 20th centuries, as Golla (1995) so beautifully described. Among these were other well-known collections

³This observation is echoed by several language archive directors working in North America in Wasson et al. 2016.

⁴See [https://en.wikipedia.org/wiki/Breath_of_Life_\(language_restoration_workshops\)](https://en.wikipedia.org/wiki/Breath_of_Life_(language_restoration_workshops))

⁵Ahtna, Incorporated is one of thirteen Alaska Native Regional Corporations created under the Alaska Native Claims Settlement Act of 1971.

including the American Philosophical Society, The National Anthropological Archives, The University of California at Berkeley, and The Jacobs Collection at the University of Washington. All of these had long been important for collecting and maintaining the analog anthropological record, but even in 1995 Golla was able to see important changes on the horizon. In particular Golla noted that digital databases would allow for the decentralization of materials so that they could be accessed in satellite locations. He mainly mentioned other university- or college-based “research centers” like the then-incipient Native American Language Center at UC Davis, but also tribal-sponsored collections like that of the Confederated Tribes of the Coos, Siuslaw, and Lower Umpqua, kept “at their tribal offices in Coos Bay, Oregon, a quite thorough archive of the published and unpublished documentation of their traditional languages (all of which are now extinct), including copies of sound recordings made by linguistic fieldworkers” (Golla 1995:157).

When Himmelmann was writing his treatise, the world was on the cusp of a digital revolution that would bring new procedures for digitizing and sharing information. Universities could buy digitization equipment and storage space for converting their collections relatively cheaply. Individual language workers could now produce born-digital language documentation to deposit in increasingly-digital language repositories. Access to digital information was also becoming easier; in North America in particular, the ramp-up to high speed internet in many remote locations was relatively quick in comparison with some regions of the world.

The onset of the digital era represents a turning point for language documentation archives. Universities and other non-Indigenous institutions had previously assumed the role of being the only qualified keeper-of-the-record, where interested audiences would be allowed to come visit materials.. This well-meaning—but in hindsight, imbalanced—dynamic was not immediately reversed by the early years of the digital revolution, but institutional archives soon began to investigate better ways to provide access to records. The DATA project represented one archive’s attempt to take advantage of these digital advances for the sake of getting language information into the hands of Alaska Native people. No longer did an interested Dena’ina person need to drive ten hours from Kenai, or charter a plane from Nondalton, to access the language materials at the ANLC in Fairbanks. One needed simply visit the qenaga.org website from a browser window in one’s own living room or at the local school.

In recent years there have even been some steps toward archives better acknowledging the needs of the archive user (Shepard 2016, Wasson et al. 2016) who is more likely to be a member of an Indigenous language community than a non-Indigenous linguist. Expectations about expertise and authority have shifted along with these changes, away from prioritizing expertise held by specialists in institutional archives, and centering the authority and expertise of Indigenous communities in determining the stewardship of records of their knowledge.

We believe this is a welcome change. As Wasson and colleagues (2016) have observed,

[...] archives are constructed within a paradigm of Western scientific concepts and assumptions [...] This includes curation practices that serve as a form of control or even suppression when decisions as to what is put in or kept out of an archive are made solely by archivists and linguists, rather than by members of the communities whose language data are being placed in the archive[...] (Wasson et al. 2016: 650)

In kind, some institutional archives in North America are now shifting to a support role, rather than positioning themselves as the sole body capable of maintaining collections. One example of this is the *Indigitization* project⁶ at the University of British Columbia, which provides training and equipment to Indigenous communities in the preservation of knowledge, but does not demand that the resultant digital resources be lodged with the university. The C'ek'aedi Hwnax Ahtna language archive has a similar arrangement with the University of Alaska Fairbanks: UAF provides long-term “grey storage” backup of all the digital language materials at no charge, and also turns over all decisions about access to the Ahtna Heritage Foundation. Another notable effort is the FirstVoices project⁷ of the First Peoples' Cultural Council,⁸ which provides tools for documentation, archiving, and dissemination of Indigenous languages.

Along with accessing archival documentation, community-based researchers in Native North America are themselves a large and growing constituency practicing language documentation and description. There have always been, in the hundreds of communities across Native North America, community-based scholars whose mission is the carrying-forward of their knowledges and traditions, and there have always been community-based experts in language use. But recent discourses around Indigenous and decolonizing research methods (cf. Kovach 2009; Smith 2012; Wilson 2008, *inter alia*) contribute to and reflect a paradigm shift in the academy which has increased recognition of research on language and culture generated beyond the ivory tower. University-based researchers from Indigenous American communities have impacted, shifted, and expanded traditional university-bound notions of what language documentation is and should be, and for whom (Begay 2017; Cranmer 2015; Jacobs 2011; Leonard 2007; Lukaniec 2018; Rosborough 2012; *inter alia*). Universities in Canada and the United States are increasingly supportive of community-engaged research, encouraging partnerships, and seeking to welcome Indigenous researchers whether they choose to work within or outside of universities. Many funders now request letters of support from community partners, and/or a Memorandum of Understanding indicating some degree of community support for a proposed project (cf. Government of Canada 2012). Universities and funders also increasingly recognize the complexity of community-engaged work, and are learning to shape expectations for project results accordingly.⁹

Community-external linguists have also been deeply impacted by their long-term working relationships with community partners, through which they have gained broader perspectives on how Indigenous communities follow protocol, set research priorities, identify research questions, frame research processes, and define key concepts such as ‘language’, ‘culture’ and ‘territory’ (cf. Czaykowska-Higgins 2009; Leonard 2017; Sapién & Thornes 2017). The past decade has seen a profusion of literature related to linguistic research which explores in depth concepts of collaboration, partnership, and appropriate models for research in community contexts (Amery 2009; Crippin & Robinson 2009; Czaykowska-Higgins 2009; Leonard & Haynes 2010; Leonard 2017; Shaw 2001; Whaley 2011; *inter alia*). Institutional growth has followed suit. The American Indian Language Development Institute at University of Arizona (AILDI ca. 1978) and Canadian Indigenous Languages and Literacy Institute at University of Alberta (CILLDI ca. 2000) are university-based programs supporting Indigenous-centered capacity building focused

⁶<http://www.indigitization.ca/>

⁷<https://www.firstvoices.com/>

⁸<http://www.fpcc.ca/>

⁹See Whaley 2011 for some examples of such complexities.

on language reclamation. The biennial Institute on Collaborative Language Research (CoLang), initiated in 2008, as well as the Stabilizing Indigenous Languages Conference (SILS ca. 1993) International Conference on Language Documentation & Conservation and the journal *Language Documentation & Conservation* also provide vital venues for knowledge exchange between academic and community experts.

This influence is evident in evolving best practices for the design of language documentation projects. The ‘lone-wolf’ approach to fieldwork asked Western-trained linguists and anthropologists to generate a research proposal, determine research questions, lay out a methodology, and seek funding in isolation without consultation with the community in question. Outside researchers might arrive in a community with a language documentation plan that had received input from their advisors on campus but did not reflect community protocols, goals, and intentions. In a broader global context, there are situations in which such an approach may still be the most appropriate model (see Crippen & Robinson 2009, 2011 and Bower & Warner 2017 for a discussion of this), but in the North American context, linguists working in this way risk replicating the extractive dynamics of colonial policies which took both language and land from Indigenous communities. Language work is time consuming. The time of Elders who are willing to share their language is particularly precious; their knowledge is key to efforts to reclaim and revitalize language and culture. For this reason, many linguists, whether outsiders or community members, feel a strong ethical imperative to ensure that their research is guided by community intentions and contributes to community priorities for language reclamation.

In North America, documentation projects may be initiated by communities with revitalization in mind; others may result with outreach from university researchers to communities, or descend from previous relationships. In any case, relationship-building lays the groundwork for an emergent and iterative approach to project design (Hermes et al. 2012). Research goals are set in response to community intentions, articulated through a process of consultation; work-in-progress is shared at key points with community stakeholders, allowing for feedback. Outcomes may evolve as the project develops. Czaykowska-Higgins points out that “...in community-based research it is often the case that *the process itself is a result*” (Czaykowska-Higgins 2009: 43).

Documentary linguists have often felt the imperative to ‘capture’ as much ‘data’ as possible in order to provide a full picture of an endangered ‘language’, but language documentation projects emerging from consultative partnerships and reflecting community-based goals of language reclamation may not prioritize comprehensiveness, nor seek to create a representative sample of a language. In fact, they may not consider language to be an object at all, but rather relate to it as a living being; a medium through which the world is experienced; a conduit for communication to ancestors (cf. Leonard 2017; Hermes et al. 2012).

Despite a vast diversity of community contexts in which languages are being documented for purposes of reclamation, several themes emerge as shared among many North American communities engaged in such projects. Especially in contexts of urgent necessity, community-based researchers may prioritize recording texts which are most likely to be useful for teachers and learners over texts which are widely representative. Certain types of language that have traditionally been valued by academic researchers, such as monologic formal speeches or sacred stories, may be less likely to receive attention. Both academic and community-based researchers note the high value of audio and video documentation of conversational speech, interaction, questions and answers

(Mithun 2001; Sammons & Rosenblum in prep) to language reclamation efforts, as well as to understanding the dynamic structures of interaction. Many projects also emphasize the importance of documenting ‘everyday language’ and daily routines of life at home. Documentation for reclamation may also need to respond to the specific features of a given language structure in order to provide useful material for teaching and learning, such as the semantically-rich and morphologically complex ‘beautiful words’ of Kwak’wala (Kell et al. 2011; Rosborough et al. 2017) which are treasured by language learners but are only produced in certain documentation contexts (Rosenblum 2015).

Finally, a shared theme among many projects is the documentation of place-based knowledges. Documenting ecological knowledges of territory occupies a special privilege for many communities, with concrete relevance and associated sensitivities. These themes motivate language research for communities continuing to live in their ancestral homelands (Cruz 2017), as well as for those who have been relocated or removed from their traditional territories. The Myaamia, whose homelands are south of the Great Lakes but now live in diaspora extending from Ohio and Indiana through Kansas to Oklahoma (and indeed, around the world), researched moon phases, plant names, and seasonal descriptions and recover traditional ecological knowledge held in the language contained in archival manuscripts in order to create a lunar calendar which is now in wide use (Voros 2009; Wigram 2009).

The effort to document place-based knowledge can be part of a larger community movement to connect younger generations with their homelands, to reclaim knowledge of those places in the language which belongs to it, to reoccupy territories and to heal from past trauma.

At the same time, in Canada, cases concerning the territorial rights of First Nations are a crucial ongoing piece of the Indigenous response to colonial occupation (cf. *Delgamuukw v. British Columbia* 1997; *Tsilhqot’in Nation v. British Columbia* 2014; Nair 2018). Many First Nations in British Columbia are actively negotiating treaty settlements. As a result, when language documentation involves knowledge of territory, harvesting practices, and traditional use and occupation, it is not only highly valued, it can involve information which is privileged and may need to be protected for various reasons. Language documentation projects must be able to plan for and accommodate such concerns; for this reason and many others, open-access requirements on data recorded within such a project may need to be flexible and responsive to these community needs.

In North America, the need for language documentation and revitalization is inextricable from the history that led to language loss. In reflecting on varying relationships to language documentation, Hermes, Bang & Marin note that for Indigenous communities, “the language revitalization movement is passionate, political, and deeply personal, particularly for many Native people who are acutely aware that the federal government’s attempted genocide was the direct cause of Indigenous language loss” (Hermes, Bang & Marin 2012). Leonard defines language reclamation as ‘a larger effort by a community to claim its right to speak a language and to set associated goals in response to community needs and perspectives’...Reclamation is thus a type of decolonization. Rather than exhibiting a top-down model in which goals such as grammatical fluency or intergenerational transmission are assigned, it begins with community histories and contemporary needs, which are determined by community agents, and uses this background as a basis to design and develop language work” (Leonard 2017: 19). In this framing, language documentation has the potential—and many would say, the

responsibility—to contribute to decolonial and anticolonial projects within and beyond Indigenous communities. Such work is emotionally heavy, and as outsiders we two authors recognize that we can never fully understand the burden of that history as it weighs on our research partners.

Given this opportunity to reflect in print, we two authors have observed that, over the past two decades, a shared and consultative approach to project design has inevitably led us and those around us to expand concepts of what research is, how it is approached, and what it should produce. But there are still plenty of steps to be taken. Looking ahead, we both hope that the institutions within which we work can continue to expand definitions of ‘language’, including allowing for multiple definitions to co-exist; to adjust the scope of what is considered ‘documentation’; and to allow the research process and its products to be determined by teams of experts, crucially involving members of speech communities. We are optimistic that two decades from now, language workers in North America will be able to look back to today and be proud of how language documentation has evolved to reflect the priorities of the communities it is intended to serve.

References

- Administration for Native Americans. n.d. Esther Martinez initiative: Preserving the heart of our cultures. (<https://www.acf.hhs.gov/ana/resource/emi-preserving-the-heart-of-our-cultures>) (Accessed June 27, 2018)
- Amery, Rob. 2009. Phoenix or relic? Documentation of languages with revitalization in mind. *Language Documentation & Conservation* 3. 138–48.
- Begay, Kayla. 2017. *Wailaki grammar*. PhD dissertation, University of California, Berkeley.
- Berez, Andrea L., Taña Finnesand & Karen Linnell. 2012. C'ek'aedi Hwnax, the Ahtna Regional Linguistic and Ethnographic Archive. *Language Documentation & Conservation* 6. 237–252.
- Berez, Andrea L. 2011. Directional reference, discourse, and landscape in Ahtna. PhD dissertation, University of California, Santa Barbara.
- Bowern, Claire & Natasha Warner. 2015. 'Lone wolves' and collaboration: A reply to Crippen & Robinson (2013). *Language Documentation & Conservation* 9. 59–85.
- Boynton, Jessica, Steven Moran, Anthony Aristar & Helen Aristar-Dry. 2006. E-MELD and the School of Best Practices: An ongoing community effort. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 87–98. Sydney: University of Sydney Press.
- Briggs, Charles & Richard Bauman. 1999. 'The foundation of all future researches': Franz Boas, George Hunt, Native American texts, and the construction of modernity. *American Quarterly* 51. 479–528.
- Carpenter, Brian. 2018. CNAIR Stories: The Kwakwaka'wakw Manuscripts of George Hunt. American Philosophical Society (blog). (<https://www.amphilsoc.org/blog/cnair-stories-kwakwakawakw-manuscripts-george-hunt>) (Accessed May 1, 2018)
- Cope, Lida & Susan D. Penfield. 2011. 'Applied linguist needed': Cross-disciplinary networking for revitalization and education in endangered language contexts. *Language and Education* 25. 267–71.
- Cranmer, Laura. 2015. Reclaiming Kwak'wala through co-constructing Gwanti'lakw's vision. PhD dissertation, University of British Columbia.
- Crippen, James A. & Laura C. Robinson. 2013. In defense of the lone wolf: Collaboration in language documentation. *Language Documentation & Conservation* 7. 123–35.
- Cruz, Emiliana. 2017. Documenting landscape knowledge in Eastern Chatino: Narratives of fieldwork in San Juan Quiahije. *Anthropological Linguistics* 59. 205–231.
- Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation* 3. 15–50.
- Davis, Jennifer L. 2013. Learning to 'talk Indian': Ethnolinguistic identity and language revitalization in the Chickasaw Renaissance. PhD dissertation, University of Colorado.
- De Korne, Haley & Wesley Y. Leonard. 2017. Reclaiming languages: Contesting and decolonising 'language endangerment' from the ground up. *Language Documentation and Description* 14. 5–14.
- Community University Engagement Support Fund. UBC Office of Community Engagement. (<http://communityengagement.ubc.ca/scholarly-resources/cues/>) (Accessed June 28, 2018)


- Golla, Victor. 1995. The records of American Indian linguistics. In Sydel Silverman & Nancy J. Parezo (eds.), *Preserving the anthropological record*, 143–157. New York: Wenner-Gren Foundation for Anthropological Research.
- Government of Canada, Social Sciences and Humanities Research Council. 2012. Social Sciences and Humanities Research Council Guidelines for the Merit Review of Indigenous Research. May 11, 2012. (http://www.sshrc-crsh.gc.ca/funding-financement/merit_review-evaluation_du_merite/guidelines_research-lignes_directrices_recherche-eng.aspx)
- Harding, Anna, Barbara Harper, Dave Stone, Catherine O’Neill, Patricia Berger, Stuart Harris & Jamie Donatuto. 2012. Conducting research with tribal communities: Sovereignty, ethics, and data-sharing Issues. *Environmental Health Perspectives* 120. 6–10.
- Hermes, Mary. 2012. Indigenous language revitalization and documentation in the United States: Collaboration despite colonialism. *Language and Linguistics Compass* 6. 131–42.
- Hermes, Mary, Megan Bang & Ananda Marin. 2012. Designing Indigenous language revitalization. *Harvard Educational Review* 82. 381–402.
- Hermes, Mary & Mel M. Engman. 2017. Resounding the clarion call: Indigenous language learners and documentation. *Language Documentation and Description* 14. 59–87.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–95.
- Himmelman, Nikolaus P.. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelman & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin, New York: Mouton de Gruyter.
- Hunt, George. Manuscript in the language of the Kwakiutl Indians of Vancouver Island, with interlinear translations [preface by Franz Boas, reviser. New York, 5 Nov. 1913], Call Number: X898 K979 H912, Rare Book & Manuscript Library, Columbia University in the City of New York.
- Holton, Gary, Andrea L. Berez & Sadie Williams. 2006. Building the Dena’ina language archive. In Laurel Evelyn Dyson, Max Hendricks & Stephen Grant (eds.), *Information technology and indigenous people*, 205–209. Hershey: Idea Group.
- Holton, Gary. 2014. Mediating language documentation. *Language Documentation and Description* 12. 37–52.
- Jacobs, Peter William. 2011. Control in Skwxwu7mesh. PhD dissertation, University of British Columbia.
- Kell, Sarah. 2014. Polysynthetic language structures and their role in pedagogy and curriculum for BC Indigenous languages. Victoria, BC: Aboriginal Education Team, BC Ministry of Education. (https://www2.gov.bc.ca/assets/gov/education/administration/kindergarten-to-grade-12/aboriginal-education/research/polysynthetic_language.pdf)
- Kovach, Margaret Elizabeth. 2000. *Indigenous methodologies: Characteristics, conversations, and contexts*. Toronto: University of Toronto Press.
- Lamer, Antonio, Gérard V. La Forest, Claire L’Heureux-Dubé, John Sopinka, Peter deCarteret Cory, Beverly McLachlin & John C. Major. 1997. *Delgamuukw v. British Columbia*. 3 SCR 1010. Supreme Court.
- Leonard, Wesley Y. 2007. Miami language reclamation in the home: A case study. PhD dissertation: University of California, Berkeley.

- Leonard, Wesley Y. 2017. Producing language reclamation by decolonising 'language.' *Language Documentation and Description* 14. 15–36.
- Leonard, Wesley Y. & Erin Haynes. 2010. Making 'collaboration' collaborative: An examination of perspectives that frame linguistic field research. *Language Documentation & Conservation* 4. 268–293.
- Linn, Mary S. 2014. Living archives: A community-based language archive model. *Language Documentation and Description* 12. 53–67.
- Lukaniec, Megan. 2018. A grammar of Wendat. PhD dissertation, University of California, Santa Barbara.
- McCarty, Teresa. 2003. Revitalising Indigenous languages in homogenising times. *Comparative Education, Special Number (27): Indigenous Education: New Possibilities, Ongoing Constraints* 39. 147–63.
- McCarty, Teresa. 2017. Commentary: Beyond endangerment in Indigenous language reclamation. *Language Documentation and Description* 14. 176–84.
- McCarty, Teresa & Tiffany Lee. 2014. Critical culturally sustaining/revitalizing pedagogy and Indigenous education sovereignty. *Harvard Educational Review* 84 (1). 101–24.
- McLachlin, Beverly, Louis LeBel, Rosalie Silberman Abella, Marshall Rothstein, Thomas Albert Cromwell, Michael J. Moldaver, Andromache Karakatsanis & Richard Wagner. 2014. *Tsilhqot'in Nation v. British Columbia* SCC 44, [2014] 2 SCR 257. Supreme Court.
- Mithun, Marianne. 2001. Who shapes the record: The speaker and the linguist. In Newman, Paul & Martha Ratliff (eds.), *Linguistic Fieldwork*, 34–54. Cambridge: Cambridge University Press.
- Nair, Roshini. 2018. B.C. First Nation files title claim to challenge fish farms in traditional territory. CBC News. June 5, 2018. (<https://www.cbc.ca/news/canada/british-columbia/b-c-first-nation-files-title-claim-to-challenge-fish-farms-in-traditional-territory-1.4691589>)
- Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5. 187–207.
- Robinson, Laura & James Crippen. 2015. Collaboration: A reply to Bower & Warner's reply. *Language Documentation & Conservation* 9. 86–88.
- Rosborough, Patricia Christine. 2012. *K̓angəxtola Sewn-on-Top: Kwak'wala revitalization and being Indigenous*. PhD dissertation, University of British Columbia.
- Rosborough, Trish, chuutsqa Layla Rorick & Suzanne Urbanczyk. 2017. Beautiful words: Enriching and Indigenous Kwak'wala revitalization through understandings of linguistic structure. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes* 73. 425–37.
- Rosenblum, Daisy. 2015. A grammar of space in Kwakwala. PhD dissertation, University of California, Santa Barbara.
- Rosenblum, Daisy & Olivia Sammons. In prep. Documenting multimodal interaction: Workflows, data management, and archiving. In *Documenting Conversation. Journal of Language Documentation & Conservation Special Publication*. Honolulu: University of Hawai'i Press.
- Rouvier, Ruth. 2017. The role of Elder speakers in language revitalisation. *Language Documentation and Description* 14. 88–110.
- Sapién, Racquel-María & Tim Thornes. 2017. Losing a vital voice: Grief and language work. *Language Documentation & Conservation* 11. 256–74.

- Shepard, Michael Alvarez. 2016. The value-added language archive: Increasing cultural compatibility for Native American communities. *Language Documentation & Conservation* 10. 458–479.
- Smith, Linda Tuhiwai. 1999. *Decolonizing methodologies: Research and Indigenous peoples*. London, New York, Dunedin, N.Z.: Zed Books; University of Otago Press.
- Stebbins, Tonya. 2012. On being a linguist and doing linguistics: Negotiating ideology through performativity. *Language Documentation & Conservation* 6. 292–317.
- Stebbins, Tonya N. & Birgit Hellwig. 2010. Principles and practicalities of corpus design in language retrieval: Issues in the digitization of the Beynon Corpus of early twentieth-century Sm'algayax materials. *Language Documentation & Conservation* 4. 34–59.
- Voros, Craig Matthew. 2009. Myaamia calendar project phase II: Lunar calendar calibration. MA thesis, Miami University.
- Wasson, Christina, Gary Holton & Heather S. Roth. 2016. Bringing user-centered design to the field of language archives. *Language Documentation & Conservation* 10. 641–681.
- Whaley, Lindsay J. 2011. Some ways to endanger an endangered language project. *Language and Education* 25. 339–48.
- Wigren, Laura. 2009. Myaamia lunar calendar project phase II: Using new technology to build mutual learning. MA thesis, Miami University.
- Wilson, Shawn. 2008. *Research is ceremony: Indigenous research methods*. Halifax: Fernwood Pub.
- Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*. Cambridge: Cambridge University Press.
- Woodbury, Anthony C.. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description* 12: 19–36.


Daisy Rosenblum

daisy.rosenblum@ubc.ca

 orcid.org/0000-0002-6803-7956

Andrea L. Berez-Kroeker

andrea.berez@hawaii.edu

 orcid.org/0000-0001-8782-515X