Reflections on documentary linguistics

Bradley McDonnell, Andrea Berez-Kroeker, Gary Holton
University of Hawai‘i at Mānoa
1890 East-West Rd., Honolulu, HI 96822, USA
[mcdon@hawaii.edu, aberez@hawaii.edu, holton@hawaii.edu]

Abstract

In order to better understand the advances in the field of documentary linguistics over the past two decades, we invited 33 disciplinary experts from around the world to reflect on key issues in the field. These issues 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. The contributors prepared short vignettes on topics ranging from “ethics” to “training” to “utilization of documentation” to “(de-)colonialism” and beyond. Some topics have been hotly debated over the past two decades, while others have emerged more recently. The contributors trace the development of these key issues both in terms of progress and possible missteps, reflecting on their past importance, current relevance and future directions in language documentation. In this paper we provide a summary of issues discussed and attempt a synthesis of the individual reflections, noting both successes and failures and outlining a vision for the next two decades of language documentation.

Looking back

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. Of course, Himmelmann was far from 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 (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). But 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 grammar-writer, 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 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 this objection misses the point. While the two activities may go hand in hand, the products suffer very different fates, with the former having much more potential to be broadly impactful to the field. Woodbury defines documentary linguistics as “the creation, annotation, preservation and dissemination of transparent records of a language” (2011: 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 doesn’t mean that we need to stop doing description. And it doesn’t 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 it 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 subfield of documentary linguistics, and in some sense Himmelmann (1998) gave birth to this field. Now 20 years later we can ask: how has this new field evolved?

To answer this question, we invited 33 experts from across the world to write short (~3000 word) reflections on key issues in documentary linguistics, with the goal of collecting them into an edited volume, to be published in late 2018 (McDonnell et al. 2018). The issues addressed by the contributors 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. The contributors trace the development of these key issues both in terms of progress and possible missteps, reflecting on their past importance, current relevance and future directions in language documentation.

We asked each contributor to reflect on a particular issue in light of (i) how the issue was originally raised; (ii) how it has developed in the field; and (iii) in what directions the issue would be most fruitfully taken. While the original mandate was to reflect on the evolution of the field in the 20 years since Himmelmann (1998), most 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 33 vignettes provide not only reflections on where we have been but also a glimpse of where the field might be headed.

While the volume is still in preparation, the 2018 Foundation for Endangered Languages Conference, with its special theme “20 years of language documentation” seemed an apt opportunity for an interim report on the volume. In what follows we summarize some of the notable observations made by the volume contributors in their draft manuscripts. Due to space limitations we are unable to share observations from all of the contributors, nor can we fully represent the many insights provided by the contributors. For that readers will have to await the final publication of the volume, which will be made available as an open-access publication through the journal Language Documentation & Conservation. Finally, while we have endeavored to faithfully represent the views of the individual contributors, those contributors have not had the opportunity to review or comment on this article. Any errors of representation or fact remain our responsibility and should not be blamed on the volume contributors.

Reflections

Here we summarize the reflections of several of the volume contributors. Bear in mind that these vignettes are not meant to be encyclopedic or comprehensive reviews of the literature but are rather intended as personal reflections on how the field of documentary linguistics has evolved with respect to a number of crucial issues. Not all contributors share the same view of documentary linguistics, and hence not all contributors see the history of the field in the same way or set the same priorities for the future of documentary linguistics. To us as editors, it is precisely this diversity which lends power to the various statements. Taken together, these various vignettes give a fuller picture of where the field has been and where it is going.

Ethics

Ewa Czykowska-Higgins begins her reflections by noting that ethical concerns have been deeply embedded within the modern conceptualization of language documentation from the outset. That said, Czykowska-Higgins acknowledges that there remains a debate about the role of collaboration in language documentation and the degree to which collaboration should be considered an ethical requirement. In recent years this debate has been prominent in the literature. Key to resolving this debate is the recognition that every language documentation situation is different, and hence we must expect that collaboration will take many different forms. If we allow space for different models of collaboration, it is easier to see collaboration as an integral part of documentary linguistics. Moving forward, Czykowska-Higgins sees a need for further evolution of language documentation practice to incorporate de-colonizing methodologies, potentially transforming the way language documentation research is done.

Language community training

Colleen Fitzgerald reminds us that “training between linguists and communities is mutual, and in fact, that such a binary distinction between linguist and community member is a false dichotomy.” Most documentary linguists would likely acknowledge that they have learned more from language communities than the other way around. In discussing the ethics of field work, Rice reflects that “all the speakers with whom I have worked have been teachers” (Rice 2006: 142). Fitzgerald’s conception of language community training builds on that mutual notion of teaching and learning which is at the heart of the language documentation enterprise.

As a way of justifying the value of language community training Fitzgerald offers a top-ten list of benefits which accrue when community members involved in language documentation receive linguistic training. Her list is worth citing in its entirety:

1. The documentation (linguistic and otherwise) is richer.
2. Scientific findings are stronger and more complete.
3. Community reclamation activities are better supported.
4. Increases the potential that activities and collaborations better address concerns … like injustice.
5. This is leading to a better understanding of language revitalization and its outcomes.
6. Language documentation and revitalization training has been extended to all continents.
7. Local, grassroots community drive is essential for sustaining efforts to sustain language documentation and revitalization.
8. Indigenous expertise, ways of knowing and epistemologies can lead to a more complete understanding of language and linguistic phenomena.
9. … a distinction between linguist and community member cannot be assumed.
10. An increase in diversity among linguists.

Diversity of participation

In reflecting on the diversity of participation in language documentation, I Wayan Arka draws on his experience in Indonesia, a country rich in languages but relatively poor in linguists and language documenters. Few Indonesians have applied for major language documentation grants, let alone been awarded such grants. Arka observes that in spite of recent efforts, there remain many barriers to community member involvement, including especially a lack of motivation and a lack of leadership.

Arka explains that the sporadic and temporary nature of documentation projects make it difficult for community members to commit to language documentation work,

1 nflrc.hawaii.edu/ldc

2 Since the volume has not yet been published, we are unable to cite page numbers for contributor quotes. Readers are advised to cite contributor quotes from the published volume.
since they must also be concerned with day-to-day issues of food, housing, and basic needs. Unfortunately, academic linguists are not always well-versed in the ways of dealing with these sociological challenges, and non-local researchers may lack understanding of the local culture filters underlying speakers’ motivation to participate in language documentation. Recruiting specialists from other disciplines, such as anthropology and educational psychology, may help overcome these barriers and thus build capacity in the community. We wholeheartedly agree with Arka’s position that the success of a language documentation project ought to be measured not just by the quality of the products (archival deposits) created but also by the extent to which the project increases community awareness of the significance of documentation. This view entails an increased focus on capacity building in order to increase diversity of participation in language documentation.

(De-)colonialism

Although much of the rhetoric of documentary linguistics stresses the importance of community perspectives, the fact remains that for the most part the origins and practice of language documentation are rooted in colonial institutions which lie outside the control of local communities. Intentionally or not, these power structures can sometimes work against Indigenous language communities. Wesley Leonard discusses some of the ways this occurs. Most notable is the distinction between documentation and revitalization / reclamation—a distinction which is at best artificial and at worst detrimental to many language communities.

Leonard concludes by offering some potential “interventions” which can facilitate decolonial approaches to language documentation. These include:

- engagement with the issue of defining “language”
- recognition that speakers are not abstract entities but rather “full people with kinship networks, occupations, responsibilities, needs, hopes, and intellectual contributions that go beyond their linguistic knowledge”
- recognition of the prescriptive nature of documentation

Public awareness

As Mary Linn cogently 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 endangered languages published in outlets such as BBC and the New York Times. Ignoring the outright racist and xenophobic comments, Linn discerns three main types of negative attitudes toward endangered and minority languages: (i) a belief that a common language creates less conflict; (ii) a kind of Social Darwinism which views language loss as inevitable; and (iii) a general apathy toward issues that don’t have immediate effect on the commenter.

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 the majority language community 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.

The scope of language documentation

Over the past two decades language documentation has emerged as an ad-hoc response to increased awareness of the endangered languages “crisis.” The race to catalog languages, develop recording and archiving standards, and implement funding schemes has left little time to reflect on the scope of language documentation as a (sub) discipline of its own. Moving forward, Jeff Good 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 we are omitting 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.3

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 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” (1988: 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.

3 It should also be recognized that the term language documentation has two distinct senses, one referring to a discipline or activity, the other referring to a product of a documentation effort. The latter sense is a count noun, and can be pluralized (cf. Seifart 2008; A. C. Woodbury 2014). The alternative term documentary linguistics is unambiguously synonymous only with the former sense.
Descriptive/documentary adequacy

Sonja Riesberg explores some of the negative emotional reactions to Himmelmann (1998) which have emerged over the past two decades, suggesting that these are largely triggered by two factors: (i) the separation of documentation from description; and (ii) the way Himmelmann challenges how linguists deal with primary data, both in terms of “defining a language” and in terms of replicability and accountability. Some of the misperceptions regarding the separation of documentation and description have already been addressed by Himmelmann in subsequent publications (e.g., Himmelmann 2012). As Riesberg points out, separating documentation from description does not imply that the latter is unnecessary or less useful. Rather, description is part of documentation, but not its purpose. Somewhat paradoxically, the success of the documentary linguistics enterprise has led in many cases to the loss of the descriptive component of language documentation.

The second emotional reaction noted by Riesberg involves the issue reproducibility and transparency in language work. The emergence of documentary linguistics over the past two decades is largely responsible for nudging linguists toward a greater appreciation for transparency and the need to back up their analyses with data. Twenty years ago, linguistics was still largely an armchair field, with argumentation supported by intuition. Where primary data were cited, these were generally not accessible. Documentary linguistics, with its focus on building publicly-accessible and repurpose corpora, was thus naturally seen as anathema by some. But the increase in the number of language documentation collections available is evidence that these attitudes are changing. Linguists—and not just documentary linguists—are beginning to image a more data-centric science of language (cf. Berez-Kroeker et al. 2018).

Why local cultural meanings matter

Modern documentary linguistics was from the start envisioned as an interdisciplinary effort, yet as Lise Dobrin and Mark Sicoli point out, the anthropological perspective is often lacking. Even some of our most basic discourses about linguistic methodologies are uninformed by community contexts. Consider the following questions which arise in any language documentation project:

- who counts as a native speaker?
- how do speaker numbers get tabulated?
- should language recordings be archived?
- what is language documentation?
- what counts as “success” in language reclamation?

None of these questions can be answered in the absolute but instead require understanding of local cultural perspectives. To take just the first of these, in situation 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. They go on to make a case for the use of participant observation methodology as a way to uncover local meanings in language documentation fieldwork.

Documentary corpora

Sally Rice argues cogently for the use of corpus tools and corpus analyses in language documentation. The creation of corpora has always been at the heart of the language documentation enterprise, but despite increasing attention to corpora for language documentation, many misunderstandings remain as to just what is meant by a language documentation corpus. Many field linguists still refer to their text collection or their FLEEx/ELAN database as a “corpus,” yet these materials may or may not be actually usable with standard corpus linguistics tools and methods. In contrast, Rice argues that a true language corpus is more than just an archive of available material. The key distinction has to do with what sort of research can be accomplished using a language corpus.

Corpus tools can help to achieve many of the core goals originally proposed for documentary linguistics, allowing language data to speak for themselves independent of description and analysis. Rice concludes that documentary linguistics would be well-served by greater integration of corpus linguistics methods.

Linguistic fieldwork

In reflecting on linguistic fieldwork over the past two decades, Claire Bowern examines three questions:

1. Have fieldwork methodologies changed?
2. Have the products of documentation changed?
3. Has the academic culture surrounding language documentation changed?

In response to the first question Bowern notes that linguists do a lot more recording now (though this is perhaps more due to the emergence of structured elicitation tools than to the articulation of the description versus documentation distinction). This shift is of course greatly facilitated by the emergence of digital recording and transcription technologies. There is a much greater emphasis on conversational and naturalistic data, and an increased explicitness regarding the nature of data and methodology. Alongside this change is a calming of the debate about the value of particular types of data (e.g., conversational vs. elicited). Fieldworkers increasingly recognize the need to collect a variety of data types.

Regarding the products of documentation, Bowern notes especially the increased focus on pedagogical materials and the embedding of archiving into the documentation process. On the other hand, many of the products of documentation remain essentially unchanged. For example, most digital grammars differ little from print publications and fail to take advantage of the potential for linking to documentation corpora (as suggested for example by Thieberger 2009).

Regarding academic culture Bowern is more pessimistic. In spite of the fact that more documentation is now conducted by or in collaboration with native speakers, the
academic discourse surrounding language documentation still presents barriers to community participation (Davis 2017). That said, the distinction between linguist and community member has been blurred, and students now receive much better training in fieldwork, ethics, and data management.

**Linguistic fieldwork across the globe**

There have clearly been many changes in the way we do linguistic fieldwork over the past twenty years, but the practice of fieldwork varies significantly across the globe. To get a better sense of this variation we sought contributions from authors working in eleven different regions of the world: Kalahari Basin; Australia; the Pacific; Canada; México; Amazonia; the Chaco; the Southern Cone of South America; India; Eastern Indonesia; and the Caucasus. While this list cannot hope to cover the all of the types of fieldwork situations which exist, it at least gives some sense of the extent to which linguistic, social, and political situations vary. Below we briefly discuss four of these fieldwork vignettes in more detail.

**Fieldwork in Eastern Indonesia** Yusuf Sawaki and I Wayan Arka note that in the Eastern Indonesian context, managing expectations is crucial to project success. This may require knowledge and expertise beyond linguistics—such as the recognition of the need to avoid the word “project” (Indonesian proyek), since it carries certain expectation of income and development. Moreover, active participation of the local community is especially essential in Eastern Indonesia.

**Fieldwork in México** Gabriela Pérez Báez identifies four critical topics for fieldwork in México:

- participation of speakers
- consent of speakers
- interests of speakers
- needs of speakers

One of the greatest barriers to building capacity in the region remains the reliance on English in training programs. Community access to documentation materials also remains a challenge, both owing to lack of local repositories and English-based interfaces at major international digital language archives.

**Fieldwork in India** Shobhana Chelliah identifies a unique challenge to fieldwork in India deriving from contesting linguistic ideologies. On the one hand, value is placed on linguistic diversity, but on the other hand value is placed on large, religiously-relevant languages. India has long privileged its larger languages, so it can be difficult to allocate resources to documenting smaller minority languages. Moreover, Indian linguistics has traditionally focused on description rather than documentation. Changing this focus will require a cultural shift that rewards the archiving of primary data.

**Fieldwork in Ghana** Felix Ameka writes about the history of linguistic research on Ghanaian languages from the missionary grammars and dictionaries of the mid-19th century forward through today. Of special note to Ameka is the recent rise of “insider” and “outsider” researchers; that is, Ghanaians who research their own languages and Ghanaians who research other languages of Ghana respectively. Ameka supports this development, since it provides opportunities for Ghanaians to gain skills in language work. He advocates for continuing the development of models in linguistic research that allow for more documentation to be undertaken by insiders.

**Overcoming the transcription challenge**

Reflecting on the progress of documentary linguistics over the past twenty years, Nikolaus Himmelmann identifies a significant remaining challenge, which he calls the transcription challenge. This problem goes far beyond the practical issues of transcribing a large amount of collected data, i.e., the transcription bottleneck. The transcription challenge is not about finding new ways to amplify the transcription process through forced alignment and speech-to-text. Rather, the transcription challenge is about recognizing transcription as a “language making” process which affects language ecology. Here Himmelmann echoes Good’s call for greater theorization of documentary linguistics, for despite its important role in language documentation, transcription remains critically undertheorized and understudied. Language documentation needs to reach a “better understanding of the transcription process itself and its relevance for linguistic theory.”

Although much work remains to be done in order to develop such a theory of transcription, Himmelmann offers some initial takeaways. First, we need to theorize language documentation as a (sub)discipline to:

- include a more nuanced view of “language” which recognizes variation and social context
- better understand the transcription process and its effects on the researcher and speaker communities
- make use of ethnographic methodologies

Second, language work with communities needs to:

- incorporate Indigenous conceptualizations of documentation
- include better training opportunities and more capacity development
- build language reclamation into the documentation cycle

To the extent documentary linguistics has concerned itself with transcription, it has mostly focused on technical issues such as tools (ELAN, FLEx, Transcriber, etc.) and standards (.eaf, IGT, etc.). 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.

**Looking forward**

The contributions to McDonnell et al. (2018), including those discussed above, bring a variety of perspectives on the evolution of documentary linguistics over the past
twenty years. This is to be expected given the personal nature of a "reflections" piece. Yet, a number of common threads run through many of the pieces, suggesting some shared perspectives on the progress of documentary linguistics, and perhaps more importantly, its future directions. If we were to name one big takeaway from these contributions it is this: after twenty years of focus on mechanics and technical standards, documentary linguistics is now poised to transition into a genuine subfield of linguistics, with its own theoretical underpinnings.

This point is made explicitly in Jeff Good’s contribution, but it is also implicit in the contributions of Himmelmann, S. Rice, Dobrin & Sicoli, Drude, Linn, Fitzgerald, and many others. For the past twenty years (or more) documentary linguists have been in the trenches recording languages and making sure that the infrastructure in place to ensure that those recordings endure. Documentary linguists survived the analog-to-digital transition and emerged with robust standards for digital language archiving. They developed funding schemes dedicated to endangered language archiving, with generally accepted sets of outcomes for documentary projects. And they developed dedicated training programs for both academic and community linguists, helping to build capacity in the field.

Now it is time to come up for air and assert documentary linguistics as something more than a mere data-gathering activity (though data remain central to the field) but as a fully theorized subfield of linguistics. This will involve moving beyond the “how to” focus of many of the language documentation handbooks which have appeared over the last two decades and instead focusing on the what and the why. Many of the concepts which documentary linguistics has taken for granted actually deserve more theoretical reinforcement. notions of transcription, corpus, culture, bilingualism, and even language itself remain poorly understood within the field of documentary linguistics.

One final point. While we explicitly asked contributors to focus on issues relating to language documentation, the issue of language revitalization and reclamation—or what is sometimes called language conservation—came up repeatedly in the contributions. There is increasing awareness that while documentation and conservation may be separable at a conceptual level, it is not possible to separate them in practice. Any proposed theorization of documentary linguistics as a field must necessarily articulate a relationship to revitalization and reclamation.

Documentary linguistics has come a long way in the past twenty years, but as Austin puts it, “much work remains to be done” (Austin 2016: 164). We remain grateful to the various contributors for more clearly articulating a path forward.

Acknowledgements

We gratefully acknowledge the volume contributors for sharing their knowledge, both as authors and as internal reviewers: Wayan Arka, Alexandre Arkipov, Christine Beier, Claire Bowern, Shobhana Chelliah, Ewa Czykowska-Higgins, Lise Dobrin, Sebastian Drude, Patience Epps, Penfield Fitzgerald, Alex François, Lauren Gawne, Lucia Golluscio, Jeff Good, Thomas Gültemann, Nikolaus Himmelmann, Wesley Leonard, Mary Linn, Marisa Malvestitti, Susan Penfield, Gabriela Perez-Baez, Sally Rice, Keren Rice, Sonya Riesberg, Daisy Rosenblum, Yusuf Sawaki, Sebastian Schnell, Mandana Seyfeddinipur, Mark Siculo, Ruth Singer, Manana Tandashvili, Nicholas Thieberger, Alejandra Vidal, Fernando Zúñiga.

References


