2

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. 13–21
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24804

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

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

ISBN: 978-0-9973295-3-7

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 *gíing'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.

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 Himmelmann & 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.

Slotta, 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