Adapting methods of language documentation to multilingual settings

Jeff Good

University at Buffalo

jcgood@buffalo.edu

Adapting methods of language documentation to multilingual settings

Jeff Good

University at Buffalo

jcgood@buffalo.edu

Abstract

Commonly recommended methods for documenting endangered languages are built around the assumption that a given documentary project will focus on a single language rather than a multilingual ecology. This hinders the potential usability of documentary materials for the study of language contact. Research in domains such as ethnography and sociolinguistics has developed conceptual and analytical tools for understanding patterns of multilingual usage, but the insights of such work have yet to be translated into concrete recommendations for enhancements to documentary practice. This paper considers how standard documentary approaches can be adapted to multilingual contexts with respect to activities such as the collection of metadata, the use of ethnographic methods, and the recording and annotation of naturalistic multilingual discourse. A particular focus of the discussion are ways in which documentary projects can create better records of multilingual practices even if these are not the focus of the work.

Keywords: language documentation, multilingualism, linguistic repertoires, methods

1 Documenting lexicogrammatical codes or language ecologies?

Over the last several decades, work on the documentation of endangered languages has resulted in the development of a full-fledged subdiscipline known as documentary linguistics (see, e.g., Himmelmann 1998; 2006).[1] Documentary linguistics has, to this point, focused primarily on the methods required to document, and in some cases maintain or revitalize, individual languages on their own terms, most typically from the perspective of language as a lexicogrammatical code rather than, for example, as a socially embedded practice (see, e.g., Woodbury 2011; Good

[1] I would like to thank participants at the Linguistic Society of America satellite workshop *Multilingualism, Contact, and Documenting Endangered Languages* held on January 6, 2019, for their feedback on the presentation on which this paper is based as well as Pierpaolo Di Carlo, the special issue editors, and three anonymous reviewers for comments on an earlier version of this paper. The work underlying the research results presented here has been supported by the National Science Foundation under Award Nos. BCS-1360763 and BCS-1761639.

2018). In effect, this means that documenting language contact tends to be backgrounded in most work in language documentation—or even consciously set aside when projects specifically limit their focus to a version of a language that is ostensibly free of phenomena such as borrowing or codeswitching, even if this is not reflective of the actual linguistic practices of a community or results in the construction of a code which did not previously exist. (See Dobrin & Berson 2011 for a critique of this practice.)

A reluctance to engage with certain kinds of language contact phenomena in language documentation projects is understandable. If an indigenous language has become endangered due to contact with a major language of a settler society (e.g., English or Spanish in the Americas), a community may specifically want to document as much as possible about how that language was used before that contact situation developed. However, in many other cases, such an approach is harder to justify. For instance, there are parts of the world where multilingualism traditionally has not merely been an incidental fact of life but has played a central role in defining and maintaining local social structures (see, e.g., Di Carlo et al. 2019 for a review of relevant work on a selection of rural African communities). In such settings, a failure to attend to the multilingual behaviors of individuals in a documentary project will paint a distorted picture of the communicative practices of their communities. It will also make it extraordinarily difficult to recover the detailed patterns of language use from the documentary record that form the most overt reflexes of language contact and, thereby, limit our ability to study language contact phenomena in settings where endangered languages are still used today but may not be in the near future.

The purpose of this paper is to consider what adaptations to standard documentary methodology are required when a project seeks to include multilingual behaviors as part of the record of the linguistic practices of a community. The perspective offered here is strongly informed by work that has been focused on understanding patterns of multilingualism in a linguistically diverse rural region of the Cameroonian Grassfields known as Lower Fungom (see Di Carlo & Good 2020 for a number of papers presenting results of this work). However, there are similar patterns of so-called small-scale multilingualism in need of documentation found elsewhere (see, e.g., Campbell & Grondona 2010 and Epps 2018 on South America, Kroskrity 2018 on North America, Rumsey 2018 and Singer 2018 on Australia, and Lüpke 2016; Pakendorf et al.

2021 for global perspectives).[2] Therefore, even if many of the examples discussed here are drawn from African contexts, the paper's general points should be applicable to communities throughout the world, though they will likely require local adaptations.

In order to provide more context for the rest of the paper, a brief overview of the currently dominant approach to language documentation and how it relates to the proposals to be made in the rest of this paper is provided in Section 2.

2 Expanding on standard approaches to language documentation

Himmelmann (1998) is generally understood to be the first work to clearly articulate the basic principles of language documentation, as distinct from language description (see also McDonnell et al. 2018). A succinct characterization of language documentation can be found in Woodbury (2011: 159), who defines it as "the creation, annotation, preservation, and dissemination of transparent records of a language". As indicated by this definition, work in this area has been built around the idea that the object of study of documentary work is specific languages, rather than, for instance, verbal repertoires, understood as "the totality of linguistic forms regularly employed in the course of socially significant interaction" (Gumperz 1964: 137), whether these are drawn from a single language or a multilingual language ecology (see Mühlhäusler 1992 for relevant discussion of the latter notion in the present context). Given that documentary linguistics emerged out of concerns regarding global patterns of language endangerment (Himmelmann 1998: 161), most documentary projects further focus on "ancestral" codes (Woodbury 2005), i.e., lexicogrammatical patterns that are taken to represent the way a language was used before the impact of recent patterns of language contact and shift.

The actual products of documentary work typically include audio and video recordings that are selected to exemplify the use of a specific language, the creation of time-aligned annotations for these recordings, and the curation of metadata that describes the records collected about the language (Thieberger & Berez 2012). These products are well suited to support the creation of

[2] The term small-scale multilingualism refers to patterns of multilingualism found in small-scale societies rather than qualifying the nature of the multilingualism itself. See Lüpke (2017: 276–277) for commentary on how patterns of small-scale multilingualism have not been fully appreciated due to colonialist assumptions about the relationship between language and identity in small-scale societies, as well as due to incorrect ideas that view multilingualism as a primarily urban phenomenon.

descriptive resources about a language such as dictionaries, grammars, and annotated texts, as well as community-oriented outputs such as standard orthographies and teaching materials. At the same time, this process of selection limits the scope of what is documented considerably, and, in particular, results in the suppression of multilingual behaviors in the resulting record, whether deliberately or inadvertently (see Section 5.2). This, in turn, reduces the range of potential uses of a documentary corpus, for instance making it difficult to study the ways that different lexicogrammatical codes are deployed across social contexts within a community or how an individual's linguistic repertoire influences their patterns of language use—topics of clear interest to the study of language contact, among other areas. Moreover, this approach can result in a record of a community's patterns of language use that is not reflective of actual patterns at the time of data collection, which, if not intended, is far from ideal. Given that documentary linguistics arose in a context where language endangerment was in focus, the fact that the initial decades of its development emphasized models and methods for documenting individual languages, rather than multilingual ecologies, is understandable. At the same time, it would clearly be desirable to consider how work in this area can be adapted to settings where multilingualism is an important part of a community's communicative practices.

The central proposal of this paper is that the documentation of multilingual ecologies can be achieved by adapting dominant approaches to documentation in four ways. The first involves placing greater emphasis on understanding the life histories, social characteristics, and linguistic repertoires of the individuals whose language use is recorded as part of a documentary project (Section 3). The second involves working with members of language communities to understand how they categorize the language varieties that they have knowledge of and how this relates to other local cultural categories (Section 4). With this information, it becomes possible to develop a plan for capturing a range of multilingual interactions that are likely to be representative of actual multilingual language use within a given community as well as a scheme for annotating them (Section 5). Finally, certain kinds of linguistic interactions that language users engage in should be recognized as being especially useful for providing further context to multilingual linguistic practices (Section 6). Taken together, the recommended adaptations would result in an expansion of the activities associated with language documentation that can

complement the kinds of work that dominate this area of investigation today and which would result in documentary records that, among other things, could be more readily used to study language contact.

This paper is organized with the documentary practitioner in mind—whether or not such an individual is a community member or an outside researcher—and, as such, the discussion in each section is intended to be relatively self-contained and focused on a related set of documentary activities. Where possible, work that exemplifies how the various recommendations have been put in practice is cited for readers interested in getting further details about the steps needed to make use of them for data collection and analysis. The overall set of recommendations made in this paper are summarized in Table 2 in Section 7. As will be clear below, some of the recommendations are made more tentatively than others, and this paper will hopefully prompt a longer term discussion of how to address the complex question of how to document multilingual ecologies.

3 Documenting language users

3.1 Understanding people's linguistic lives

Structural grammatical patterns of the sort that are the focus of traditional descriptive linguistic investigation are idealized as being properties of a language rather than emanating from individual-level practices. This is, of course, a simplification, but, for many kinds of linguistic concerns, e.g., determining the phoneme inventory of a language or whether or not it makes use of a case system, it is a useful one. In documentary projects, basic information about the individuals whose language use is recorded is generally tracked for purposes of long-term archiving and to make sure that these creators are properly acknowledged. However, in part because of the idealization just mentioned above, the specific life histories of these language users, and how those histories may be relevant for understanding their patterns of language use, are not generally seen as specific targets of investigation.

Sociolinguistic studies of variation in understudied languages are exceptions to this (see, e.g., the papers in Stanford & Preston 2009), but these are still atypical in the discipline and tend to emphasize how specific language user characteristics can help account for observed varia-

tion within a language rather than treating the linguistic behavior of individuals themselves as fundamental to the research. From the perspective of Eckert's (2012) delineation of variationist sociolinguistic research into three "waves", these would probably be best considered second-wave studies, with analyses built on a model of how locally salient social categories relate to language use, rather than emphasizing only macrosociological categories such as sex or age, as is typical of first-wave approaches.

Documenting multilingualism, by contrast, requires an approach that places the life history, status in the local social system, and linguistic repertoire of an individual language user at the center of the research.[3] This falls more in line in with third-wave sociolinguistic studies, which put language user agency into much greater focus in the study of variation (see Eckert 2016 for general discussion and Di Carlo et al. 2019; Di Carlo this issue for further discussion in a documentary context). In particular, multilingual usage can only be fully analyzed and understood in light of the linguistic knowledge that individuals have at their disposal, and it is not possible to understand their choices without knowing their linguistic repertoires and their relationship to the other individuals that they are interacting with at a particular time.

To underscore this point, consider the fragment of dialog in Figure 1, drawn from Tabe (2020: 131). It was recorded in the village of Ossing, in the Southwest Region of Cameroon. Ossing is associated with two local languages, Kenyang [ken; keny1279] and Ejagham [etu; ejag1239].[4] There are three participants in the dialog: Mayok and Ntui, both natives of Ossing, and "Messié" (an adaptation of the French word *monsieur*), a Cameroonian who is not native to Ossing. Ossing is in the so-called anglophone region of Cameroon, where Cameroon Pidgin English [wes; came1254] dominates as the local lingua franca. Messié is from the so-called francophone region, where French is widely used. Messié is a male admirer of Mayok, and she wants to limit what he knows about her and, therefore, asks Ntui not to reveal certain aspects of her personality to him. The different languages in the dialog below are indicated using the following conventions: italics for *Kenyang*; italics and underlining for *Ejagham*; roman type for English; roman and underlining for French; bold for **Cameroon Pidgin English** (CPE), and bold

[3] As such, it has parallels to the *translanguaging* approach that has been developed to support the analysis of multilingual behaviors in research within applied linguistics (Wei 2018).

[4] When languages are introduced in this paper, they are followed by the ISO 639-3 code and glottocode (Hammarström et al. 2021) for the language.

and italics for ***Camfranglais***, a mixed variety incorporating elements of Cameroonian French and English and Cameroon Pidgin English (see Kießling 2005).

Mayok: Messié cheri, bonsoir! ***How nɔ cheri!***
"My dear sir/Messié, good evening! [French] How are you, dear? [Camfranglais]"

Messié: ça va ma cherie
"I'm fine, my dear [French]"

Mayok: [to Ntui] **longtime nɔ see! how you loss so?** *βéléé ka ká ywé ámiŋé ká nnɨk mwét ɔ kɛ ka yí á dɨ́ŋé mé*
"[to Ntui] It has been long since we saw one another, I hope all is well with you? [CPE] Don't betray my character to Messié [Ejagham], please, don't let him know me [Kenyang]."

Ntui: *ɔ ké dén mwét, ʧɔ́ŋ ŋ́ kwuti wɔ* as usual *ɔ fòntí má áfón, ɔ fiŋ ntí ma ákám* **me and you deh boh**
"You don't have to be worried, I will protect you [Kenyang] as usual [English]. You and I [CPE] are birds of a feather [literally, 'We both have many things in common in life'] so there is need to shield our poor behavior in public and in particular to strangers [Ejagham]."

Mayok: **Messié we go for bar nɔɔ**
"Messié, shall we go to the bar? [CPE]"

Messié: on pars alors **but I not have drink cherie because I sick malaria. Come we go my shine shine baby** *ŋgɔ́ré dɨ́k*
"We can go then to the bar [French]. However I will not have a drink because I am sick with malaria [Camfranglais]. Come on, let's go my beautiful lady [CPE], my beautiful queen [Kenyang]."

Figure 1: Multilingual conversation recorded in Ossing, Cameroon (Tabe 2020: 131)

It would be impossible to understand the significance of the language choices in the dialog in Figure 1 without knowing the linguistic repertoires of the participants. This is perhaps most evident in considering the turn between Mayok and Ntui which makes use of the two local languages as a means to prevent Messié from understanding what is being said. The exchange is otherwise dominated by local languages of wider communication such as French, English, and Cameroon Pidgin English, though Messié does make use of one short phrase in Kenyang to compliment Mayok at the end of the fragment. Among other things, it is crucial to know that Mayok and Ntui both have Kenyang and Ejagham in their linguistic repertoires while Messié does not, at least beyond a minimal degree. It is also important to know that Messié is from the francophone part of Cameroon. Otherwise, the use of as much French as is found in the dialog

would be hard to understand given that it records an event that took place in Ossing, which, as mentioned above, is in the anglophone region of Cameroon. For further discussion of the analysis of this fragment in terms of its structure and the social meaning of language choice, see Tabe (2020: 128–131).

A recording of a dialog like the one seen in Figure 1 would be a valuable target for a documentation project placing multilingualism at its center, or even as an example of multilingual practices within a community for a project focused on a single language. It is also clear that ensuring that it would have value as an actual documentary record of the practices of the community would require that, alongside the recording itself, rich information about the language users is also collected. I discuss possible ways of gathering this information via questionnaires in Section 3.2, along with consideration of how to interpret the validity of data derived from questionnaires in Section 3.3.

3.2 Self-characterizations of linguistic repertoires

While uncovering some aspects of linguistic identity will require extensive interaction with an individual, useful information can be obtained relatively quickly through the use of questionnaires. (See Di Carlo this issue for detailed consideration of the use of questionnaires in the documentation of multilingualism that complements the discussion presented in this section.) These questionnaires can ask an individual to report, for example, on the languages that they are able to understand and use, the contexts in which they learned those languages, and when they typically use them. They can also include questions that help situate an individual in their social network, such as asking for information about parents, spouses, and children or membership in local groups. To be maximally useful, it is important for such questionnaires to be ethnographically informed. That is, the questions must be constructed in a way that matches local understandings of language difference and language use, and they must be designed to gather information on salient social categories within the culture of the individuals who are taking part in the documentation project.

In Figure 2 and Figure 3, two parts of a questionnaire designed to gather information on patterns of multilingualism in the Lower Fungom region of Cameroon, drawn from Esene Agwara

Paternal name
Maternal name
Other names
Gender
Date of birth
Occupation
Paternal affiliation
Maternal affiliation
Spouses' provenance
Spouses' languages
Father's provenance
Father's languages
Mother's provenance
Mother's languages
Children's languages

Figure 2: Part of a questionnaire used to investigate multilingualism in Cameroon

Language name
Degree of competence
Where learned
Where used
Advantages of knowing the language
Special contexts of use (e.g. prayers, songs, invocations)

Figure 3: Questions for each language that a consultant reports knowledge of

(2013: 118–119), are presented as a means to illustrate how questionnaires could fit into a documentation project, whether or not it is focused on multilingualism specifically. The questions in Figure 2 are designed to obtain general information about an individual alongside some information about their family, while the questions in Figure 3 are repeated for each language that an individual reports knowledge of. The questionnaire items are presented in English. However, this particular questionnaire is designed as a prompt to the interviewer to cover a number of specific topics, communicating with the interviewee using whatever language and strategy is most appropriate and feasible.

Some of the questions in Figure 2 and Figure 3 would be more or less universally applicable, such as those intended to gather information about an individual's date of birth or occupation. Others would still be broadly applicable in most of the world, but their precise formulation is aligned with the goal of documenting multilingualism, as is seen in the use of the plural *languages* throughout, rather than the singular *language*, in the questions in Figure 2 in order to

not implicitly suggest a monolingual norm to a consultant. This particular questionnaire is very strongly "code"-based, in the sense that it is organized around asking individuals what languages they and their family members know, though other kinds of configurations are conceivable, too, such as a domain-based questionnaire which might ask how individuals would communicate in different social contexts.

Simply using a questionnaire like the one in Figure 2 and Figure 3 is not enough to ensure that useful data is collected. It is also important to employ it in a way which maximizes comparability of answers across surveyed individuals. This has been facilitated for this questionnaire by accepting and encouraging answers that make use of locally salient categories (see also Section 4). For instance, while Figure 3 uses the term *language* informally to help guide the interview, the data collected actually focuses on locally salient named lects, which may or may not correspond to languages in the scholarly linguistic sense (Esene Agwara 2020: 189). Since, in this setting, the named lects all refer to linguistically salient varieties that are known to outside researchers, it is not hard to associate them with recognized languages following the current scholarly classification.

A questionnaire like the one presented in Figure 2 and Figure 3 allows for the collection of much richer individual metadata than is possible with more standard approaches, and this metadata can be used, in turn, to analyze multilingual exchanges among individuals. At a minimum it helps establish what languages any given set of individuals may have in common, which sets a potential "baseline" through which to analyze their language choices. When such metadata is aggregated, it can also be a key part of the documentation of the language practices of a community by allowing for the analysis of patterns within the linguistic repertoires of its members (see Esene Agwara 2020 for an example).

A potential concern regarding questionnaire data is that, since questionnaires are necessarily designed around some set of theoretical assumptions, there are likely to be dissociations between data derived through the use of questionnaires and what would be revealed through observation of language in use or what might be discovered through long-term interaction with members of a community (see also Section 3.3). A good example of the significance of information discovered in the latter way can be found in the discussion of the life history of Hélène Coly presented

by Lüpke & Storch (2013: 24–28) (see Di Carlo this issue: §2.2.2 for additional discussion). She is a woman from the Lower Casamance region of Senegal and underwent a ritual to change key aspects of her identity, including her primary linguistic identity, after she had difficulty having children. Such a pattern would be unlikely to be revealed through any kind of general questionnaire. However, since sociolinguistic questionnaires for multilingual documentation should be understood as evolving tools, they could be adapted to cover the discovery of newly discovered links between language and identity as the understanding of the sociolinguistic features of a community evolves.[5]

Before concluding the discussion in this section, it would be worthwhile to contrast the kinds of individual metadata that can be collected through a sociolinguistic questionnaire from the sort of individual metadata associated with widely used metadata standards for linguistic resources. The OLAC standard (Simons & Bird 2008), for instance, is not designed to capture metadata about individuals in any systematic way since it assumes that the role of metadata is to describe resources themselves. The more expansive IMDI standard (ISLE Metadata Initiative 2003; Broeder & Wittenburg 2006) does provide for the systematic encoding of metadata with individuals, treating them as *actors*. Individuals are still primarily seen as contributors to resources. However, there is a relatively rich set of categories that can be specified for individuals such as their age, sex, primary language, education, role within the event being recorded, and social role among the individuals participating in the recorded event (ISLE Metadata Initiative 2003: 17–19). Nevertheless, these categories can not provide nearly the same coverage as a questionnaire like the one presented in Figure 2 and Figure 3. In a project focused on a single language, information of the kind that can be encoded via the IMDI standard may provide more or less sufficient context for interpreting patterns of language use in the recording. For a project emphasizing documentation in multilingual settings or one interested in gathering data relevant to understanding patterns of language contact in a given community, it is clear that more extensive individual metadata is needed.

[5] In addition, in some communities, individuals may choose not to fully reveal their linguistic repertoire, in particular to a researcher with whom they do not have any kind of established relationship. This is a further reason why a questionnaire cannot be seen as a substitute for longer-term interactions which can yield unexpected and important information.

Fully administering a sociolinguistic questionnaire of the sort described in Section 3.2 can take a fair amount of time—perhaps an hour or more—especially if an individual has knowledge of many languages. Entering the collected data into some kind of standardized format will also take additional time. This, therefore, represents an additional burden on a researcher compared to standard approaches employed today, in particular if a project is not focusing on the documentation of patterns of multilingualism. If the task is limited to principal consultants, for instance in a project primarily focused on documenting a specific language, it will probably be quite manageable. Moreover, it has the benefits of enhancing the documentary record both by providing a richer range of information about the consultants and also of the kind of linguistic knowledge held within the community itself. It also has the less tangible, but still important, benefit of providing a means for the researcher to come to a better understanding of the linguistic lives of their consultants than traditional metadata gathering practices support.

3.3 Aligning self-characterizations with other data

Self-reported information on an individual's linguistic knowledge and patterns of language use cannot be expected to perfectly align with actual patterns of use. Among other reasons, this is because self-reported information will inevitably be mediated by local language ideologies (see also Section 6). An example in a documentary context is provided by patterns of language use in an in-law avoidance register of Datooga, a Nilotic language of Tanzania, as described by Mitchell (2015). This register, referred to in Datooga using the label *gíing'áwêakshòoda*, requires married women to avoid the names of a large set of their in-laws as well as words that sound like those names. It is connected to a larger set of avoidance behaviors that center on fathers-in-law. For instance, a daughter-in-law should avoid any physical contact with her father-in-law (Mitchell 2015: 125). The use of the avoidance register alongside the non-avoidance register in Datooga communities does not present a canonical instance of multilingualism. However, as argued by Di Carlo & Neba (2020), the study of distinctive registers within varieties that are categorized as the same "language" in scholarly linguistic sources can reveal patterns that are also clearly relevant to the study of multilingualism, which makes consideration of such work relevant here.

The register is described by women as being used at all times (Mitchell 2015: 112), reflecting the fact that its use is a salient part of a complex of behavioral patterns associated with proper behavior for daughters-in-law. Actual usage is not as strict as what is reported. Mitchell (2015: 208–213) found, for instance, significant variation in the extent to which certain lexical items (as opposed to names) were avoided on the basis of an examination of the speech of two women. On the whole, avoidance of lexical items that sound similar to names which must be avoided appears to be adhered to less closely than the avoidance of names themselves (Mitchell 2015: 200). However one might choose to analyze this pattern, this dissociation between reported usage and observed usage is clearly something that should be documented since it is part of the linguistic culture of the Datooga community.

It is easy to see how the results from this study of *gíing'áwêakshòoda* can be analogized to more canonical multilingual settings. For instance, an individual may report that they use a given language fluently, but they are never observed to actually use it, even in contexts where it would seem appropriate. This might suggest that there is some social prestige associated with knowledge of the language causing them to report greater competence in it than they actually have. Alternatively, an individual might be observed to use a language that they did not report knowledge of, raising questions as to why they did not report this knowledge in a survey (see Evans (2001) for detailed discussion of relevant examples in an Australian context).

Despite the fact that naturalistic observational data is clearly of high value, it will often not be feasible to gather the full range of such data required to assess how well self-reported data aligns with actual usage, especially when one is dealing with reported patterns of usage across multiple languages and more than a handful of members of a language community. More active elicitation-based and task-based strategies can, in principle, be employed to assist with the interpretation of individuals' self-reports, though there does not appear to be much work in this area.

Mba & Nsen Tem (2020) consider this issue in their development of methods to assess multilingualism among residents of Lower Fungom, Cameroon, the same area that was the focus of the work of Esene Agwara (2020), discussed in Section 3.2 (see also Di Carlo this issue: §4.2.1). Three different methods were used to assess an individual's linguistic competence. The first was

an adaptation of a method known as Recorded Text Testing, as developed by Casad (1974: 3–50), which has long been used to test the comprehension that a user of one dialect of a language has of other dialects within a dialect complex.[6] In a multilingual context, this same approach can be used to assess the extent to which self-reported information from individuals regarding their passive competence of a given language aligns with their ability to understand recorded texts in that language.[7] In Lower Fungom, Mba & Nsen Tem (2020: 215–216) found that individuals' reported degrees of passive competence were in line with the results that emerged from RTT.

Mba & Nsen Tem (2020) also developed two tests to assess active competence—that is the ability to speak or sign a language (in addition to understanding it). The first is quite straightforward, though it does not appear to have been systematically explored in the literature. This is to elicit wordlists from all the languages for which an individual reports active competence. The collected wordlists can then be compared with a reference wordlist collected from an individual who would uncontroversially be considered a fully competent user of that language. While documentary work generally privileges data from so-called "native speakers", in this approach, the object of study is not a "language" (i.e., a lexicogrammatical code) but, rather, the linguistic repertoire of a given individual, thus reversing the standard relationship between a consultant and a language in linguistic work.[8] That is, rather than working with a consultant to get data about a language, reference data from a language is used to get a better understanding of the consultant's linguistic knowledge. The use of this method, as described in Mba & Nsen Tem (2020: 217–219), produced interesting variation in results. For instance, individuals sometimes reported not knowing what the translational equivalent was for a word in one of the languages that they claimed knowledge of, as might be expected, while in other cases they produced a word that deviated strongly from the reference word, suggesting that they sometimes overestimated their knowledge. In other cases, they were able to produce the expected root but with inflectional morphology that did not match what was found in forms collected from the reference speakers.

[6] See Yoder (2017) for a recent overview and appraisal of RTT.

[7] The term *passive competence* is used here to refer to cases where an individual is able to understand the use of language even if they cannot speak or sign it themselves.

[8] It may not be possible to determine which individuals might qualify as "native speakers" in advance when conducting documentary work in highly multilingual contexts, in particular when the languages used within a community are not yet well described. However, this does not prevent use of the relevant patterns of usage from being documented, and later analysis may help reveal which language users are most likely belong to this category. On the notion of *language* in a documentary context, see Good (2018).

While the nature of this study did not allow for strong inferences based on the data, it showed clear promise as a means of assessing active competence.

In a similar vein, Mba & Nsen Tem (2020) also explored the use of visual stimuli to assess active competence. Individuals were presented with images depicting scenes relevant to daily life in Cameroon. They were then recorded commenting on them in a particular language. These recordings were played back to individuals who were considered to have native-level competence in those languages and who judged the language use of the individual responding to the stimuli. Mba & Nsen Tem (2020: 216) found that people's ability to speak the relevant languages was in line with their reported levels of competence, or even higher than the reported level in many cases.

Gathering self-reported information via questionnaires or similar instruments, as discussed in Section 3.2, can be somewhat time consuming but generally not prohibitively so for all but the shortest documentary projects. By contrast, the kinds of assessment described in this section are much more time consuming and would be difficult to incorporate into most projects. The strategies for assessing active competence, in particular, can be burdensome to implement due to the fact that they rely on working not only with one's primary consultants but also individuals who can judge the linguistic abilities of those consultants, given that, in a highly multilingual settings, even the most experienced researcher is unlikely to have the necessary levels of knowledge of all of the languages used within a given community to do this work on their own. This includes cases where a member of the community is also a researcher, since they may not have the required degree of knowledge of all of the languages under investigation and, therefore, will likely need to make use of additional judges in at least some cases.

A more realistic strategy is to undertake the assessment of multilingual competences for a manageable sample of individuals as a means to help interpret and "calibrate" the data gathered via questionnaires. If this kind of work were done more widely, and in different parts of the world, the results from these studies could perhaps be used to help interpret self-reported data on linguistic knowledge in nearby communities on the assumption that cultural conventions for reporting levels of linguistic competence may be broadly similar within a given linguistic area. Admittedly, however, this last point must be considered speculative without dedicated research.

4 Documenting local linguistic categories

When to classify a set of varieties as a "language", as opposed to a "dialect" or multiple distinct languages, is known to present a number of complications (see, e.g., Cysouw & Good 2013). Documentation projects focusing on a single language can generally avoid addressing this concern since it is straightforward to document the lexical and grammatical characteristics of a set of similar underdescribed varieties even when one is not able to set specific boundaries on how those varieties fit into some global classificatory scheme of the world's languages, or even a more localized classificatory scheme.

However, it is difficult not to address concerns regarding linguistic boundaries in a multilingual documentation project. For instance, in a dialog such as the one presented in Figure 1, being able to document instances of codeswitching presupposes that one can classify a given stretch of language in use as belonging to one code over another. This involves relatively well-known problems such as how to distinguish cases of code-switching from borrowing (see, e.g., Myers-Scotton 1992 for relevant discussion) as well as problems that have emerged more clearly as a result of recent efforts to put multilingual practices at the center of documentation efforts.

Watson (2019) provides a good example of work along the latter lines in her development of a framework for conceptualizing language boundaries building on work on prototype theory as developed by Rosch (1999). Prototype theory is useful for modeling complex categories whose members do not necessarily adhere to a set of definitional criteria in a straightforward way. For instance, the category encompassed by the English word *bird* would include a set of animals strongly associated with specific properties such as being able to fly, having feathers and wings, laying eggs, etc. Some members of the category would be seen by English speakers as more prototypical than others, even if other, less prototypical members would uncontroversially be seen as members of the category (e.g., a medium-sized bird that can fly such as a robin would be seen as prototypical while a penguin would not be).

The application of prototype theory to languages is not as intuitive as its application to a category like *bird*. In Watson's (2019) approach, each language is viewed as a category and specific linguistic features are seen as more or less prototypical for that language. Within a given language community, each language can be expected to also have a set of prototypical features

FEATURE	STATUS IN KUJIRERAY	STATUS IN BANJAL
Word-initial [k]	Frequent, little variation	Rare, alternates with [g]
Word-initial [g]	Rare, alternates with [k]	Frequent, little variation
Word-initial [h]	Semi-frequent, alternates with [x]	Rare, alternates with [x]
Word-initial [x]	Semi-frequent, alternates with [h]	Frequent, little variation
Word-initial [t]	Semi-frequent, no variation	Semi-frequent, no variation

Table 1: Distribution of some initial sounds in Kujireray and Banjal (Watson 2019: 153)

which differentiate it from other languages used in the community (e.g., different words, sounds, syntactic structures, etc.), which would be emblematic of that language.

On this conception, the lexicogrammatical codes of two "languages" can be analyzed as partly overlapping, implying that, when an individual uses some set of forms, they might not be clearly speaking one language over the other, but, rather, both simultaneously, in some sense. By contrast, another set of forms may be associated with only one of the two languages. Watson (2019: 146–154) illustrates this by considering the distribution of certain segments in word-initial position in two Joola languages of the Atlantic group of Niger-Congo in close contact with each other in the Lower Casamance region of Senegal, namely Kujireray [gsl; gusi1246] and Banjal [bqj; band1340].[9] A summary of the patterns she discusses is provided in Table 1.

Kujireray and Banjal are closely related varieties and, as a result, have a significant amount of grammatical overlap, a fact which is clear to both linguists and speakers (Watson 2019: 147). At the same time, each has features which are most clearly associated with one language over the other. For instance, as can be seen in Table 1, word-initial *k* is strongly associated with Kujireray, and word-initial *g* is strongly associated with Banjal. While the associations are less straightforward, word-initial *h* and *x* also have different distributions across the two languages, while word-initial *t* does not. From the perspective of prototype theory as applied to languages, word-initially, *k* could be said to be prototypical of Kujireray and *g* could be said to be prototypical of Banjal in the local linguistic space. The presence of a sound like *t* word-initially would be prototypical of both, but the fact that it is shared across the languages would mean that it would not be an emblematic feature of either of them. Some words (e.g., those not containing *k* or *g* word-initally), therefore, could be viewed by speakers as simultaneously being drawn from

[9] See Cobbinah (2020) for relevant discussion of the linguistic situation of Lower Casamance, including difficulties in drawing clear boundaries between languages.

Kujireray and Banjal when they are used, while others (e.g., those beginning with *k* or *g*) would be seen as belonging to just one of the two languages. While the example data is phonological in nature, this potential for overlap or distinctiveness, in principle, can extend through all aspects of grammar.

Whether or not one accepts Watson's (2019) approach as an appropriate analysis for the cognitive representation of differences among languages, it helps clarify the difficulties of determining what languages are being used in interaction in a multilingual context where closely related languages are in contact. In particular, if two languages share a substantial amount of vocabulary and morphology, there may be few structural linguistic cues that allow one to determine which language is being used at a given moment, and it is not even clear that this is even a sensible question in many cases.

From a documentary perspective, it may not be possible to resolve complex issues surrounding the assignment of stretches of language use to specific languages in cases where extensive individual-level multilingualism is the norm. Section 5.3 will consider the problem of annotating multilingual data, but, for this to be done in any useful way, some reasonably stable set of categories for classifying "languages" is required. Moreover, this set needs to be discoverable in a relatively easy way rather than requiring detailed analysis of large amount of data.

In the case of multilingualism in Lower Fungom, discussed above in Section 3, it has been possible to rely on local conceptions of linguistic differentiation in the early stages of analysis (see, e.g., Esene Agwara 2020: 189). In the local sociolinguistic space, varieties are associated with specific villages and residents reliably refer to varieties at the level of the village, i.e., they are named lects in the local referential space. Not all village-level varieties in Lower Fungom would be classified as distinct languages using scholarly linguistic criteria such as mutual unintelligibility. Rather, some would be considered dialects of a single language.[10]

When gathering information via sociolinguistic surveys or annotating data with the assistance of consultants, the use of the local categorization scheme is an effective way of ensuring

[10] The Mungbam varieties of Lower Fungom, as described by Lovegren (2013: 3–6), provide an example. They are associated with five villages and are probably best viewed as constituting two very closely related languages, with four villages belonging to one dialect cluster and the variety of one village being distinctive enough to constitute its own language. Locally, however, each village is described as having its own "talk", and each is associated with a clearly distinctive variety in both local and scholarly terms.

that the data is collected in a reliable and replicable way. Such data can then serve as the basis of more detailed analyses across a variety of areas, such as establishing the ideal scholarly linguistic classification of the region's varieties, examining the relationship between local ways of classifying linguistic variation and local social structures, or determining what kinds of linguistic differentiation are seen as emblematic of distinct varieties in the local sociolinguistic space, among other things. It also provides an initial framework for analyzing patterns of language contact by establishing how individuals within a community demarcate the lexicogrammatical elements that they use into what they perceive as distinct codes. The products of such work are also more readily adapted for use by community members than would be the case if scholarly linguistic categories were employed throughout the analysis, due to the fact they make use of locally recognizable classificatory schemes.

An open question is whether an approach that focuses on locally named lects is appropriate for all multilingual contexts or if the presence of consistent local naming conventions is only found in some communities. Indeed, as discussed in Di Carlo et al. (2019: §4.2–4.4), drawing on data from Cobbinah et al. (2017), multilingual language usage in Lower Casamance is not "regimented" in the same way as found in Lower Fungom (Ojong Diba 2018; 2020).[11] In Lower Casamance, extensive codeswitching can be found in natural discourse, while this is much less typical in Lower Fungom. This flexibility appears to correlate with weaker conceptual boundaries among lexicogrammatical codes in Lower Casamance, as indicated by the fact that Watson's (2019) prototype approach was specifically designed for the Lower Casamance situation. Moreover, even if a community does have established conventions for named lects, this does not mean that the names will necessarily map to lexicogrammatical codes in ways that allow them to be straightforwardly used for linguistic analysis. Understanding how local categories can relate to scholarly conceptions of "languages" would seem to require more documentary work looking at this issue and be an appropriate priority for multilingual documentary work. Such work also has consequences for models of annotation for multilingual data (see Section 5.3).

[11] Strikingly, a lack of regimentation in code use is also found in the Ossing area, as discussed in Section 3.1, despite the fact that it is quite close to Lower Fungom geographically.

5 Documenting multilingual language usage

5.1 Recording and annotating multilingual language data

This section focuses on what would typically be considered the "core" of a documentation project: Annotated recordings of naturalistic language use. In Section 5.2, consideration is given to recording multilingual events, and Section 5.3 discusses different possible strategies for annotating such recordings. As will be made clear, while achieving representative coverage of patterns of multilingual usage and doing detailed morpheme-level analysis of multilingual data would be far outside of the scope of most projects, there are steps that almost any project could take to get at least foundational data on multilingual practices within a community.

5.2 Selecting multilingual interactions

A key consideration for any documentary project is the selection of events to be recorded. In a monolingual documentation project, a significant concern is recording events that span a range of genres (see, e.g., Himmelmann 1998: 176–183). Certain kinds of studies also require that a diverse set of individuals are involved in the project so that variation within a language can be adequately documented (see Hildebrandt 2003: 381–387 for relevant observations).

For a multilingual documentation project, the above factors are relevant, but there is the additional issue of creating records representing patterns of usage of multiple languages within a community. One thing that this requires is that more attention be paid to the setting in which an event takes place than would normally be needed for a monolingual documentation project. For instance, natural conversation among family members within their home is likely to have very different patterns of language choice than natural conversation within a market setting given the different sets of actors involved. In the former setting, the dynamic would be likely to be more fixed, reflecting the fact that individuals who live together would have stable patterns of language usage with each other, whether this involves the use of one language or multiple languages.[12] In the latter setting, interactants would be more varied and less predictable, and

[12] See, for instance, the discussion of language use patterns of the highly multilingual individual, Mbang Janet, of Lower Fungom (see Section 3.2), who, despite her ability to speak more than thirteen named linguistic varieties, was reported as only speaking Buu [boe; mund1328], the language of her father, with her children (Ojong Diba 2018: 141).

language choice, as well, would be expected to be more varied (see Connell 2009 for a study of language use in a Sub-Saharan African market setting for a relevant example).

Broadly speaking, we can expect both the setting in which an event takes place and the actors involved in the event to be especially important in influencing language choice. The linguistic repertoires of the actors necessarily constrain the range of languages that will be used, and the relationship of the actors to each other can play a significant role in shaping which of the communicatively available languages are actually chosen (see, e.g., Ojong Diba (2020: 23–26) for an example where language choice is influenced by which participant in a two-way exchange is more senior). In addition, certain languages may have more typical associations with some kinds of settings than others (e.g., public vs. private, ritual vs. everyday, etc.).

Di Carlo et al. (2021: supplementary materials) describe an approach to collecting samples of multilingual language use among different groups of individuals by having a consultant wear a visible recording device during daily activities. This allows for a range of social and physical contexts of natural conversation to be recorded as the consultant moves and encounters different individuals over the course of a day. Connell (2009: 140–141) describes an alternative approach where the setting is fixed but the space is one where diverse actors regularly come together. In this case, language choice in a market was analyzed by tracking the language used in interactions between traders and customers where transactions of only a small set of traders (who were chosen on the basis of what they sold) were examined over the course of a day. This particular approach allowed for a relatively high degree of control for some actors since the selection of traders was stable over the course of the study, while also allowing many other actors to be observed as different customers interacted with the traders. These two cases should be treated simply as examples of how multilingual data can be collected. The broad question of how to ensure that a collection of multilingual exchanges is representative of the actual practices of a community is clearly in need of further research.

Adopting strategies like those just described would mean departing from common documentary practice where recording choices tend to center on specific kinds of events rather than specific people (as was the case with the methodology described in Di Carlo et al. 2021) or places (as was done in the study of Connell 2009). Adding one or two recording sessions ori-

ented towards people or places would probably be manageable for most projects and also likely to be revealing of culturally interesting linguistic patterns, even for projects not focused on multilingualism.

However, it should also be emphasized that it is possible to gather naturalistic multilingual recordings in a more passive way. This is simply to not "de-select"' them. Due to the nostalgic orientation of most language documentation projects (see Woodbury 2011: 178), language documenters often select (whether consciously or unconsciously) events that are seen as representing an earlier state of language use before recent patterns of contact impacted them, and this leads to an emphasis on monolingual language use. Avoiding this bias would not only result in the collection of recordings of multilingual language use but also produce records that more directly reflect common patterns of use within the community. These would, of course, be more valuable for studies of language contact than recordings of "unnaturalistic" monolingual language use. (See the supplementary materials of Di Carlo et al. 2021 for more concrete proposals regarding how to collect recordings of multilingual interactions.)

Moving away from an "anti-multilingual" approach can be achieved in part by merely avoiding steps that would bias recordings towards being monolingual, such as choosing events to record primarily on the basis of whether or not they are more typically associated with monolingual language use, excluding community members from recordings because they happen to regularly use other languages as part of their typical patterns of linguistic practice, or providing implicit judgments about the language choices of individuals (e.g., by asking whether a specific word they have used is a "borrowing" or not).[13] In this context, it is probably worth bearing in mind that the simple act of a researcher entering a community and saying that they want to work on "its language" may cause community members themselves to monitor their language choice in ways which are not reflective of their patterns of language use when the researcher is not present. That is, the "de-selection" of multilingual language use could begin, whether intended or not, even before a single recording is made. An alternative approach that could be taken to avoid such an outcome would be for the research to be framed in terms of understanding simply how people communicate within a given community.

[13] See Grinevald (2007: 49–51) on the topic of working with individuals with diverse patterns of language use in a documentary context.

5.3 Diversifying annotation

Two kinds of annotations are especially commonly employed in monolingual documentation projects, transcriptions and free translations. In addition, some portion of the language materials collected are generally also associated with morpheme-by-morpheme (i.e., "interlinear") glossing to facilitate grammatical analysis (see Schultze-Berndt 2006 for extensive discussion). As evidenced by resources like the TEI guidelines (TEI Consortium 2019), these only represent a small subset of the possible kinds of annotations that have been made on linguistic data, and, to the extent that they have been privileged, it is because they play an important role in supporting structural linguistic analysis.

Annotation is one of the most time-consuming tasks in language documentation. For multilingual recordings, detailed transcription and glossing may rapidly become impractical if the recordings include the use of multiple underdescribed languages. It may be reasonable in some cases for a linguist to provide detailed annotation for multilingual data involving, for instance, codeswitching between a language of wider communication, a local lingua franca, and a single underdescribed language, especially when all of them are associated with standard orthographies. Doing something similar for conversational data involving five or more languages, more than one of which is underdescribed, along the lines of what is seen in the data presented in Figure 1, will rapidly become very challenging.

However, for multilingual data, where much can be learned by observing the circumstances under which individuals switch from one language to another, it is not always necessary to annotate the data at the level of detail that is required for structural analysis. Instead, one can employ annotations that are more specific to multilingual language use, such as analyzing a given stretch of discourse as making use of one language over another. This was the method adopted by Ngué Um et al. (2020), which studied how language choice was affected by the topic of conversation among a group of women born in different rural areas in central Cameroon but who now live in the small village of Kelleng in the Littoral Region of Cameroon. A representation of the annotation strategy that they employed is presented in Figure 4, drawn from Ngué Um et al. (2020: 59).

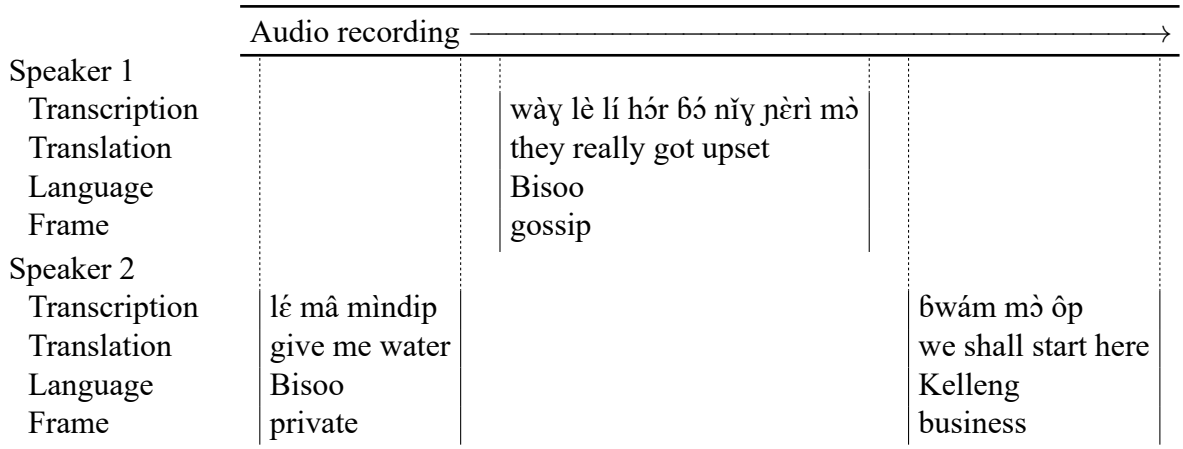

Figure 4: Multilingual annotation scheme following Ngué Um et al. (2020: 59)

Figure 4 schematizes the time-aligned annotation system adopted by Ngué Um et al. (2020) in a way comparable to what would be created by the widely-used annotation tool ELAN.[14] Annotations are linked to particular stretches of an audio recording and grouped into four categories: transcription, translation, language used (in this data, either Bisoo [bkh; biso1242] or Kelleng [btc; bati1251]), and frame of discourse (using the categories *private*, *gossip*, and *business* in this example). In addition, annotations are associated with specific speakers. A similar kind of transcription scheme, though somewhat more detailed, can be found in Cobbinah et al. (2017: 87–90). This is based on research that was also covered by the work of Watson (2019), discussed in Section 4, which considered the question of how to annotate multilingual data in underdescribed languages in what is probably the greatest level of detail found in any documentary project to date.

Perhaps due to the fact that work on documentary linguistics emerged out of discussions among linguists largely interested in descriptive and comparative work or in the revitalization of specific languages, the literature on language documentation has emphasized creating those annotations most needed for structural linguistic analysis. Consideration of multilingual data, by contrast, suggests that significant work remains to be done on both theoretical and applied concerns with respect to the range of annotations that might be valuable for different kinds of linguistic analysis. A particular area of methodological interest is the fact that annotating multilingual data will typically require much greater reliance on the knowledge of local language

[14] ELAN (https://archive.mpi.nl/tla/elan) has been created at the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands; see Brugman & Russel (2004).

users who are familiar with all of the languages being used in a given recording than is the case for projects focused on a single language, something which Di Carlo et al. (2021: 10) refer to as the challenge of relinquished control on the part of the researcher. This, of course, implicitly correlates with increased community control, a potentially positive outcome.

6 Talking about language

Collecting recordings of naturalistic language use representing diverse genres is central to language documentation. However, outside of projects focused on specific cultural domains (e.g., ritual language, traditional ecological knowledge, etc.), it is possible to vary the subject content of recordings to achieve multiple aims. On the one hand, independent of their specific subject content, the recordings can be used to support linguistic analysis, whether in structural domains, such as morphosyntax, or sociocultural domains such as the ethnography of communication (see, e.g., Hymes 1962[1971] and Michael 2011: 126–128). On the other hand, the content of these recordings can also be used to support analysis within whatever domain is covered by that content. For instance, an oral history of a language community can play an important role in understanding the community's historical relationship to other nearby communities.

In a project emphasizing multilingual language documentation, various topics could be chosen for recordings—ideally in consultation with local community members—which would yield insights into patterns of multilingual usage and, thereby, facilitate the analysis of multilingual data. Of particular value would be texts exemplifying different kinds of metapragmatic discourse (see Silverstein 1976: 48–51 and Lucy 1993: 17–18), such as the conditions under which individuals describe in their own terms why they choose to use one language over another or what their attitudes are towards users of different languages. I am not aware of any projects specifically prioritizing the collection of texts of these kinds in the languages being documented. It seems that, instead, this kind of information is more typically collected in a language of wider communication, as in the following example from Sow (2020: 143–144).[15] A masters student at the University of Ziguinchor in Senegal named Ynot describes how he learned the different

[15] In this transcription, the following conventions are used: /, //, /// for short, medium, and long pauses; *, **, *** for short, medium, and long silences.

languages of his repertoire. He reported this in French, and his description is translated into English (with language codes added).

> Je parle Mankagne parce que c'est la première langue que j'ai acquise depuis ma naissance, c'est la langue de mes parents et de mes aïeux *** le Manding, c'est dû à mon passage à Goudomp / dans le balantacounda / c'est la langue du milieu *** le Créole vient de mon passage en Guinée et des nombreux allers-retours à Bissau *** le français * l'anglais et l'espagnol sont des langues que j'ai apprises à l'école *** le wolof c'est à Ziguinchor /// Là-bas, toutes les langues sont imbriquées ou mélangées. C'est comme un melting-pot quoi.
>
> "I speak Mankanya [knf; mank1251] because it is the first language that I acquired from birth, it is the language of my parents and my ancestors *** Manding [mand1435], that is due to when I lived in Goudomp / in Balantacounda / it is the language of the environment *** Creole [kea; kabu1256] comes from when I was in Guinea and my trips back and forth from Bissau *** French [fra; stan1290] * English [eng; stan1293] and Spanish [spa; stan1288] are languages that I learned at school *** [I learned] Wolof [wol; nucl1347] at Ziguinchor /// There, all languages are interwoven or mixed. It's like a melting pot."

There is no reason, in principle, why a description like this could not have, instead, been collected with Ynot using some language other than French, such as Mankanya or Wolof, or even where multiple languages were used. A text of that kind would simultaneously provide useful information about the individual (see Section 3), an example of how metapragmatic discourse is structured in an underdescribed language (or even languages), and data that can be used to better understand the grammar of the languages used, among other things.

Investigation in the domain of language socialization is also relevant here (see Ochs & Schieffelin 1984; Schieffelin & Ochs 1986; Garrett & Baquedano-López 2002), understood as "socialization through the use of language and socialization to use language (Schieffelin & Ochs 1986: 163)."[16] Language socialization overlaps with language acquisition in being concerned

[16] Garrett & Baquedano-López (2002: 340–351) specifically discuss work on language socialization in multilingual contexts.

with how a language is acquired, with a frequent emphasis on acquisition by children. However, it is focused on the relationship between language use and social norms rather than the processes through which the characteristics of children's language use come to converge with those of adult use. I am not aware of significant work done on language socialization in monolingual documentation projects, let alone multilingual ones, though Shulist & Rice (2019: 50–52) consider how an understanding of language socialization within a given community can support work on revitalization and Hellwig & Jung's (2020) discussion of the value of child-directed speech for documentary projects is clearly relevant in this context.

Documenting language socialization requires a focus on contexts where it is especially visible, such as when children are present with caregivers or culturally common contexts where older individuals must acquire a new language in multilingual societies (e.g., when a woman moves into her husband's residence after marriage). Since there does not appear to have been significant work by documentary linguists on language socialization in monolingual contexts, it is hard to say how it would differ in multilingual contexts beyond the fact that it would be important to ensure that there is not a bias in the selection of events being recorded towards monolingual interactions (see Section 5.2).

A relevant kind of multilingual interaction that could be recorded as an instance of natural language use as well as a way of documenting language socialization can be found in Moore (2004). This study of multilingualism and language learning among individuals in a village in northern Cameroon does not specifically fall within the documentary paradigm. However, its attention to sociolinguistic details, in particular the linguistic repertoires and life histories of language users, is very much in line with points made in Section 3 regarding the need for documentary work on multilingualism to treat the characteristics of individuals as more directly relevant to the research.

In this study, Moore (2004: 135) discusses the social patterns of a highly multilingual community with respect to language learning, including a specific strategy for introducing children to new languages. From around the age of four or five, children are sent on errands where they are asked to deliver memorized messages. These can be fairly long and in a language which the child may not know at all, or at least not know well, and they provide a salient means of teach-

ing children different languages and signaling the value of learning them. Due to its overt and structured nature, such an activity would be a straightforward target for a documentary project that would simultaneously capture naturalistic multilingual data, a local language socialization strategy, and, in all likelihood, some metapragmatic discourse as well.

As pointed out by Shulist & Rice (2019: 51), processes of language socialization and, in particular, the ways that language ideologies are transmitted through them, can play an important role in maintaining the vitality of a language and in the success or failure of efforts at revitalization for older language learners. This suggests that developing recommendations on how to document language socialization effectively may have an important role in supporting revitalization and maintenance efforts. Documenting language socialization in stable multilingual contexts, in particular, is likely to yield insights into the factors that cause individuals to acquire multiple languages without resulting in language shift to a socioeconomically dominant language.

Finally, documenting both metapragmatic discourse and contexts of language socialization has clear significance for the study of language contact. Each is relevant to understanding how different languages are used within a given community and helps establish links between individual-level language attitudes and high-level effects of contact. For instance, as discussed by Ojong Diba (2020: 26–27), in Lower Fungom (see Section 3), there are general social prohibitions against code mixing in the local languages, but these are quite relaxed when it comes to Cameroon Pidgin English. This means that, if someone does not know a word in a local language, it is relatively acceptable to substitute a word from Cameroon Pidgin English but relatively unacceptable to substitute a word from another local language. It is easy to imagine how a usage pattern of this kind could result in Cameroon Pidgin English becoming the source of many borrowings into the local languages alongside more limited borrowing between local languages. Studying the socialization of this usage pattern would, therefore, be of value for understanding the mechanisms that promote borrowing from one language into another in contexts where there are many logically possible borrowing scenarios.

7 Expanding the documentary project

As indicated in Section 1, work in language documentation has typically emphasized documenting specific languages rather than taking multilingual patterns of usage as the primary object of investigation. Accordingly, generally accepted practices have yet to emerge with respect to multilingual documentation. However, it is possible to provide some concrete recommendations based on the discussion above, and, in particular, to distinguish between steps that can be taken to augment a project focused on a single language so that some information on the multilingual social reality of its users can be recorded and what would be needed for a project specifically oriented towards documenting multilingual practices.

In Table 2, a summary and partial synthesis of key points from the preceding sections is provided, presenting ways in which work on language documentation can be adapted to facilitate the documentation of multilingualism. Four potential domains of documentation are covered: (i) the multilingual repertoires of individuals (see Section 3), (ii) local categories for classifying linguistic varieties (see Section 4), (iii) multilingualism in language use (see Section 5), and (iv) metapragmatics and language socialization (see Section 6). For each domain, possible ways of extending the work associated with a monolingual documentation project are given, along with an indication of the extra effort that would be involved to support each of them.

As indicated in Table 2 some of these adaptations do not require a significant amount of extra effort and would provide immediate benefits for almost any project. These include, for instance, gathering expanded individual and contextual metadata, conducting structured interviews with people to learn about locally important categories for understanding linguistic variation, or simply not de-selecting for multilingualism when making recordings. Documenting multilingualism in some other ways would require a fundamental reworking of standard approaches rather than a simple "add-on".

The summary provided in Table 2 should be viewed only as an initial set of suggestions, and it should also be emphasized that this paper has used standard approaches to monolingual documentation as a reference point for consideration of multilingual documentation rather than trying to "reimagine" language documentation with multilingualism at its foundations (see Di Carlo

DOMAIN	EXTENSION	WORKLOAD
Individuals and their repertoires	Expanded metadata on individuals and the context of a recording	Low; immediate benefits for almost any project
	Task-based and experimental methods to assess competence of individuals across the languages of their repertoires	High; mostly for projects focused on multilingualism and contact
Local linguistic categories	Structured interviews; basic observation of community and individual life patterns; familiarization with relevant ethnographic literature	Low; immediate benefits for almost any project
	Targeted ethnographic investigation emphasizing local language ideologies; comparative analysis of lexical and grammatical features of local lects to detect emblematic features	High; mostly for projects focused on multilingualism or contact and likely to require interdisciplinary collaboration
Multilingual usage	Not suppressing multilingual usage in recordings; selecting contexts where multilingualism is likely to be found	Low; produces a more accurate record of actual language use even if only one language is the focus of analysis
	Shallow annotation of instances of multilingual language use (e.g., only of language being used)	Medium; benefits will depend on goals of specific project
	Detailed morpheme-level annotation indicating which language or languages a given morpheme can be associated with	High; mostly for projects focused on multilingualism and contact
Metapragmatics and language socialization	Recording events where language choice is discussed and behaviors connected to language socialization in multilingual settings are especially visible	Low; collected materials can be used to support structural analysis
	Targeted investigation of alignment between reported patterns of usage (ideally collected in languages being documented) and actual usage	High; best for projects focused on multilingualism and contact
	Broad investigation of language socialization activities and development of linguistic repertoires over the lifespan	High; best for projects focused on multilingualism and contact

Table 2: Summary of ways to adapt standard documentary approaches to multilingual contexts

et al. 2021 for a paper that adopts something along the lines of the latter perspective). It is inevitable that new kinds of recommendations will be needed as more work is done in this area.

I would like to conclude this paper by remarking briefly on the fact that the significance of developing a systematic set of methods for documenting multilingualism is greater than it might first appear to be. Multilingualism is now, and has been historically, a fundamental part of the linguistic lives of many language users. Multilingual usage also constitutes a salient behavioral manifestation of the abstract phenomenon of language contact. Therefore, the study of multilingual documentation is not merely about understanding the details of multilingualism in any given community—an important topic in its own right—but also about providing key data for understanding linguistic phenomena that are manifested in multilingual language use. Moreover, to the extent that language documentation tends to focus on endangered and underdescribed languages, expanding work on multilingual documentation now can play an important role in ensuring that the database for theories and models of language contact, among other areas, is properly informed by the full range of multilingualisms found in different sociocultural contexts. There is special urgency to such work given that traditional patterns of multilingualism will often be lost before languages themselves disappear as bilingualism between a local language and a language of wider communication displaces other kinds of multilingualism. That is, multilingualism, and the kinds of knowledge embedded within it, will often be even more endangered than lexicogrammatical codes themselves (see also Childs et al. 2014: 172).

References

Broeder, Daan & Peter Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies* 1. 119–132.

Brugman, Hennie & Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2065–2068. Lisbon: ELRA. http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf.

Campbell, Lyle & Verónica Grondona. 2010. Who speaks what to whom? Multilingualism and language choice in Misión La Paz. *Language in Society* 39. 617–646. https://doi.org/10.1017/S0047404510000631.

Casad, Eugene H. 1974. *Dialect intelligibility testing*. Dallas: Summer Institute of Linguistics.

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. http://hdl.handle.net/10125/24601.

Cobbinah, Alexander. 2020. An ecological approach to ethnic identity and language dynamics in a multilingual area (Lower Casamance, Senegal). In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 69–103. Lanham, MD: Lexington Books.

Cobbinah, Alexander, Abbie Hantgan, Friederike Lüpke & Rachel Watson. 2017. Carrefour des langues, carrefour des paradigmes. In Margaret Bento Michelle Auzanneau & Malory Leclère (eds.), *Espaces, mobilités et éducation plurilingues: Éclairages d'afrique ou d'ailleurs*, 79–97. Paris: Édition des Archives Contemporaines.

Connell, Bruce. 2009. Language diversity and language choice: A view from a Cameroon market. *Anthropological Linguistics* 51. 130–150.

Cysouw, Michael & Jeff Good. 2013. Languoid, doculect, and glossonym: Formalizing the notion 'language'. *Language Documentation & Conservation* 7. 331–359. http://hdl.handle.net/10125/4606.

Di Carlo, Pierpaolo. this issue. Reappraising questionnaires in the study of multilingualism: Lessons from contexts of small-scale multilingualism. *Journal of Language Contact*.

Di Carlo, Pierpaolo & Jeff Good (eds.). 2020. *African multilingualisms: Rural linguistic and cultural diversity*. Lanham, MD: Lexington Books.

Di Carlo, Pierpaolo, Jeff Good & Rachel Ojong Diba. 2019. Multilingualism in rural Africa. In *Oxford research encyclopedia of linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.227.

Di Carlo, Pierpaolo & Ayu'nwi N. Neba. 2020. The so-called royal register of Bafut within the Bafut language ecology: Language ideologies and multilingualism in the Cameroonian Grassfields. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 29–51. Lanham, MD: Lexington Books.

Di Carlo, Pierpaolo, Rachel Ojong Diba & Jeff Good. 2021. How to document multilingualism? Towards a coherent methodology: Dealing with speech data. *International Journal of Bilingualism* 25. 860–877. https://doi.org/10.1177/13670069211023144.

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: CUP.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41. 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828.

Eckert, Penelope. 2016. Third wave variationism. *Oxford Handbooks Online.* https://doi.org/10.1093/oxfordhb/9780199935345.013.27.

Epps, Patience. 2018. Contrasting linguistic ecologies: Indigenous and colonially mediated language contact in northwest Amazonia. *Language & Communication* 62. 156–169. https://doi.org/10.1016/j.langcom.2018.04.010.

Esene Agwara, Angiachi Demetris. 2013. *Multilingualism in Lower Fungom: Analyses from an ethnographically-oriented sociolinguistic survey*. Buea, Cameroon: University of Buea MA thesis.

Esene Agwara, Angiachi Demetris. 2020. What an ethnographically informed questionnaire can contribute to the understanding of traditional multilingualism research: Lessons from Lower Fungom. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 181–203. Lanham, MD: Lexington Books.

Evans, Nicholas. 2001. The last speaker is dead—long live the last speaker! In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 250–281. Cambridge: CUP.

Garrett, Paul B. & Patricia Baquedano-López. 2002. Language socialization: Reproduction and continuity, transformation and change. *Annual Review of Anthropology* 31. 339–361. https://doi.org/10.1146/annurev.anthro.31.040402.085352.

Good, Jeff. 2018. Reflections on the scope of language documentation. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 13–21. *Language Documentation & Conservation* Special Publication no. 15. http://hdl.handle.net/10125/24804.

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: OUP.

Gumperz, John J. 1964. Linguistic and social interaction in two communities. *American Anthropologist* 66. 137–153.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. *Glottolog 4.5*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642.

Hellwig, Birgit & Dagmar Jung. 2020. Child-directed language – and how it informs the documentation and description of the adult language. *Language Documentation & Conservation* 14. 188–214. http://hdl.handle.net/10125/24920.

Hildebrandt, Kristine A. 2003. *Mangage tone: Scenarios of retention and loss in two communities*. Santa Barbara, CA: University of California, Santa Barbara, PhD Dissertation.

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.

ISLE Metadata Initiative. 2003. *Part 1: Metadata elements for session descriptions (version 3.0.4)*. https://web.archive.org/web/20191031090607/https://tla.mpi.nl/wp-content/uploads/2012/06/IMDI_MetaData_3.0.4.pdf.

Kießling, Roland. 2005. 'bàk mwà mè dó'–Camfranglais in Cameroon. *Lingua Posnaniensis* 47. 87–107.

Kroskrity, Paul V. 2018. On recognizing persistence in the indigenous language ideologies of multilingualism in two Native American communities. *Language & Communication* 62. 133–144. https://doi.org/10.1016/j.langcom.2018.04.012.

Lovegren, Jesse. 2013. *Mungbam grammar*. Buffalo, NY: University at Buffalo PhD dissertation.

Lucy, John A. 1993. Reflexive language and the human disciplines. In John A. Lucy (ed.), *Reflexive language: Reported speech and metapragmatics*, 9–32. Cambridge: CUP.

Lüpke, Friederike. 2016. Uncovering small-scale multilingualism. *Critical Multilingualism Studies* 4. 35–74.

Lüpke, Friederike. 2017. African(ist) perspectives on vitality: Fluidity, small speaker numbers, and adaptive multilingualism make vibrant ecologies (response to Mufwene). *Language* 93. e275–e279.

Lüpke, Friederike & Anne Storch. 2013. *Repertoires and choices in African languages*. Berlin: De Gruyter Mouton.

Mba, Gabriel & Angela Nsen Tem. 2020. Ways to assess multilingual competence in small, unwritten languages: The case of Lower Fungom. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 205–224. Lanham, MD: Lexington Books.

McDonnell, Bradley, Gary Holton & Andrea L. Berez-Kroeker. 2018. Introduction. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 1–11. *Language Documentation & Conservation* Special Publication no. 15. http://hdl.handle.net/10125/24803.

Michael, Lev. 2011. Language and culture. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 120–140. Cambridge: CUP.

Mitchell, Alice. 2015. *Linguistic avoidance and social relations in Datooga*. Buffalo, NY: University at Buffalo PhD dissertation.

Moore, Leslie C. 2004. Multilingualism and second language acquisition in the northern Mandara Mountains. In George Echu & Samuel Gyasi Obeng (eds.), *Africa meets Europe: Language contact in West Africa*, 131–148. New York: Nova Science.

Mühlhäusler, Peter. 1992. Preserving languages or language ecologies? A top-down approach to language survival. *Oceanic Linguistics* 31. 163–180.

Myers-Scotton, Carol. 1992. Comparing codeswitching and borrowing. *Journal of Multilingual & Multicultural Development* 13. 19–39. https://doi.org/10.1080/01434632.1992.9994481.

Ngué Um, Emmanuel, Marguérite G. Makon & Célestine G. Assomo. 2020. Multilingualism as it unfolds: Language vitality in naturally occurring speech in Kelleng, a rural setting in Cameroon. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 53–67. Lanham, MD: Lexington Books.

Ochs, Elinor & Bambi B. Schieffelin. 1984. Language acquisition and socialization: Three developmental stories. In Richard A. Shweder & Robert A. LeVine (eds.), *Culture theory: Essays on mind, self, and emotion*, 276–320. Cambridge: CUP.

Ojong Diba, Rachel. 2018. *The sociolinguistic dynamics of rural multilingualism in Africa: The case of Lower Fungom*. Buea, Cameroon: University of Buea PhD thesis.

Ojong Diba, Rachel. 2020. Nuances in language use in multilingual settings: Code-switching or code regimentation in Lower Fungom? In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 15–28. Lanham, MD: Lexington Books.

Pakendorf, Brigitte, Nina Dobrushina & Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism* 25. 835–859. https://doi.org/10.1177/13670069211023137.

Rosch, Eleanor. 1999. Principles of categorization. In Eric Margolis & Stephen Laurence (eds.), *Concepts: Core readings*, 189–206. Cambridge, MA: MIT Press.

Rumsey, Alan. 2018. The sociocultural dynamics of indigenous multilingualism in northwestern Australia. *Language & Communication* 62. 91–101. https://doi.org/10.1016/j.langcom.2018.04.011.

Schieffelin, Bambi B. & Elinor Ochs. 1986. Language socialization. *Annual Review of Anthropology* 15. 163–191. https://doi.org/10.1146/annurev.an.15.100186.001115.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 213–251. Berlin: Mouton de Gruyter.

Shulist, Sarah & Faun Rice. 2019. Towards an interdisciplinary bridge between documentation and revitalization: Bringing ethnographic methods into endangered-language projects and programming. *Language Documentation & Conservation* 13. 36–62. http://hdl.handle.net/10125/24798.

Silverstein, Michael. 1976. Shifters, linguistic categories, and cultural description. In Keith H. Basso & Henry A. Selby (eds.), *Meaning in anthropology*, 11–55. Albuquerque: University of New Mexico Press.

Simons, Gary F. & Steven Bird (eds.). 2008. *OLAC metadata*. http://www.language-archives.org/OLAC/metadata-20080531.html.

Singer, Ruth. 2018. A small speech community with many small languages: The role of receptive multilingualism in supporting linguistic diversity at Warruwi Community (Australia). *Language & Communication* 62. 102–118. https://doi.org/10.1016/j.langcom.2018.05.002.

Sow, Ndiémé. 2020. Spaces and interactions in multilingual repertoire construction: A case study in an urban area of Casamance (Senegal). In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 137–154. Lanham, MD: Lexington Books.

Stanford, James N. & Dennis R. Preston (eds.). 2009. *Variation in indigenous minority languages*. Amsterdam: Benjamins.

Tabe, Florence A. E. 2020. Multilingualism in rural Africa: A case study of Ossing village in Cameroon. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 115–136. Lanham, MD: Lexington Books.

TEI Consortium. 2019. *TEI P5: Guidelines for electronic text encoding and interchange [version 3.5.0]*. TEI Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: OUP.

Watson, Rachel. 2019. Language as category: Using prototype theory to create reference points for the study of multilingual data. *Language and Cognition* 11. 125–164. 10.1017/langcog.2019.9.

Wei, Li. 2018. Translanguaging as a practical theory of language. *Applied Linguistics* 39. 9–30.

Woodbury, Anthony C. 2005. Ancestral languages and (imagined) creolisation. In Peter K. Austin (ed.), *Language documentation and description, volume 3*, 252–262. London: Hans Rausing Endangered Languages Project. http://www.elpublishing.org/PID/044.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: CUP.

Yoder, Zachariah. 2017. The reliability of recorded text test scores: Widespread inconsistent intelligibility testing in minority languages. *Journal of Multilingual and Multicultural Development* 38. 843–855. https://doi.org/10.1080/01434632.2016.1278220.