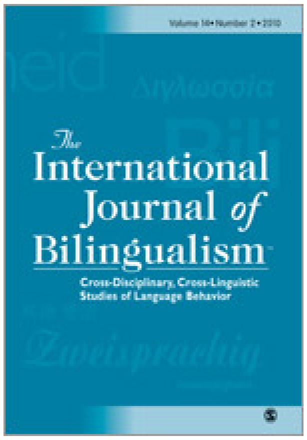

Towards a coherent methodology for the documentation of small-scale multilingualism: Dealing with speech data

Journal:	*International Journal of Bilingualism*
Manuscript ID	IJB-20-0108.R2
Manuscript Type:	Special Issue: Typology of Small-Scale Multilingualism
Keywords:	small-scale multilingualism, rural Africa, language documentation, methodology, ethnographic approaches
Abstract:	Purpose: To contribute to the establishment of a novel approach to language documentation that includes bilingual and multilingual speech data. This approach would open this domain of study to work by specialists of bilingualism and multilingualism. Approach: Within language documentation, the approach adopted in this paper exemplifies the "contemporary communicative ecology" mode of documentation. This radically differs from the "ancestral code" mode of documentation that characterizes most language documentation corpora. Within the context of multilingualism studies, this paper advocates for the inclusion of a strong ethnographic component to research on multilingualism. Data and Analysis: The data presented comes from a context characterized by small-scale multilingualism, and the analyses provided are by and large focused on uncovering aspects of local metapragmatics. Conclusions: Conducting language documentation in contexts of small-scale multilingualism requires that the adequacy of a corpus is assessed with regard to sociolinguistic, rather than only structural linguistic, requirements. The notion of sociolinguistic adequacy is discussed in detail in analytical terms and illustrated through an example taken from ongoing research led by the authors. Originality: To date, there are no existing publications reviewing in the detail provided here how the documentation of multilingual speech in contexts of small-scale multilingualism should be carried out. The contribution is highly original, in particular, for its theoretical grounding of the proposed approach.

Significance/Implications: This article can serve as a reference for those interested in methodological and theoretical concerns relating to the practice of language documentation in contexts of small-scale multilingualism across the world. It may also help clarify ways for sociolinguists to engage more closely with work on language documentation, a domain that has thus far remained primarily informed by structural linguistic approaches.

Towards a coherent methodology for the documentation of small-scale multilingualism: Dealing with speech data

Pierpaolo Di Carlo
University at Buffalo

Rachel A. Ojong Diba
University of Buea

Jeff Good
University at Buffalo

Purpose: To contribute to the establishment of a novel approach to language documentation that includes bilingual and multilingual speech data. This approach would open this domain of study to work by specialists of bilingualism and multilingualism.

Approach: Within language documentation, the approach adopted in this paper exemplifies the "contemporary communicative ecology" mode of documentation. This radically differs from the "ancestral code" mode of documentation that characterizes most language documentation corpora. Within the context of multilingualism studies, this paper advocates for the inclusion of a strong ethnographic component to research on multilingualism.

Data and Analysis: The data presented comes from a context characterized by small-scale multilingualism, and the analyses provided are by and large focused on uncovering aspects of local metapragmatics.

Conclusions: Conducting language documentation in contexts of small-scale multilingualism requires that the adequacy of a corpus is assessed with regard to sociolinguistic, rather than only structural linguistic, requirements. The notion of sociolinguistic adequacy is discussed in detail in analytical terms and illustrated through an example taken from ongoing research led by the authors.

Originality: To date, there are no existing publications reviewing in the detail provided here how the documentation of multilingual speech in contexts of small-scale multilingualism should be carried out. The contribution is highly original, in particular, for its theoretical grounding of the proposed approach.

Significance/Implications: This article can serve as a reference for those interested in methodological and theoretical concerns relating to the practice of language documentation in contexts of small-scale multilingualism across the world. It may also help clarify ways for sociolinguists to engage more closely with work on language documentation, a domain that has thus far remained primarily informed by structural linguistic approaches.

1 Small-scale multilingualism and language documentation

Arising out of concerns about language endangerment (Hale et al. 1992), and still mostly conducted in endangerment contexts, language documentation (henceforth LD) has provided a wealth of new data that has mostly supported the agenda of structural linguists (e.g., typologists, syntacticians, phonologists, etc).[1] Work within LD has been carried out mainly by adopting a mode of documentation focusing on "ancestral codes", an expression coined by Woodbury (2005, 2011) to refer to the process of creating a documentary corpus theorized (i.e., ideated) in terms of adherence to one specific language (or code) that is in danger of fading out of use in a given speech community.[2] As a result, the overwhelming majority of the LD corpora produced to date are focused on monolingual data only (see also Epps, this volume, on the need to go beyond the "ancestral code" in LD).

It is clear that this monolingual bias reflects more (Western) scholarly agendas rather than the lived reality of speech communities. For one thing, contexts of endangerment are, by definition, contexts of language shift, and this obviously implies that communities impacted by endangerment are in fact bilingual or multilingual.[3] However interesting from a scientific point of view, it is understandable that documenting multilingual speech in contexts of language shift may be seen as a paradoxical (if not masochistic) choice for a linguist working on the documentation of an endangered language with an eye on language maintenance or revitalization like many documenters are. Funding bodies also may promote a monolingual bias, even if this is unintentional.

The lived multilingual realities of an endangered language speech community may fall much more clearly within the scope of a "sensitive" documenter if we consider them from a different perspective. Recent literature has made abundantly clear that multilingualism in contexts of language endangerment is in many cases not to be ascribed only to the process of language shift, but is also engendered by indigenous societal dynamics (see, e.g., Cobbinah et al. (2017), Di Carlo (2018), Epps (2018), Singer (2018), and Rumsey (2018)).

In post-colonial contexts, which account for a large part of language endangerment contexts, if multilingualism is widespread within a certain speech community, then this is likely due to the interplay of two distinct dynamics. On the one hand, the colonial past has produced specific social and cultural dynamics relating to key areas of social life such as, for example, social hierarchy, power, and conceptions of identity, which constitute the major features of the sociolinguistic life of those societies today, especially in growing cities. This dynamic is tightly connected with the spread of ex-colonial languages and, to some degree, also of certain lingua francas—such as Cameroon Pidgin English in Anglophone Cameroon (see, e.g., Menang 2004).

On the other hand, the indigenous societal contexts continue pre-colonial dynamics which, to varying degrees, still provide ground for the reproduction of certain "endogenous" sociolinguistic processes often characterized by small-scale, relatively egalitarian social

[1] We would like to acknowledge audience members at the Typology of small-scale multilingualism conference held at the Collegium de Lyon, April 15–17, 2019, for their input on the research underlying this paper as well as the editors of the special issue and three anonymous reviewers on the manuscript itself. One reviewer, in particular, provided a series of insightful comments which we were not in a position to address in this paper, but which we hope to consider in future work. The research reported on here has been supported by NSF Awards BCS-1360763 and BCS-1761639. The order of authors reflects the fact that the first two authors contributed more or less equally to the paper, while the last author's contribution was more limited. The first author contributions focused on, in particular, the overall theoretical framing of the paper, the second author provided key data, analyses, and consideration of practical concerns related to research of the sort described here, and the third author provided general contributions to the overall structure and framing of the paper and also contributed to the text itself.

[2] In this paper, we focus our consideration on communities characterized primarily by the use of spoken languages, rather than sign languages. We leave open here the extent to which the arguments in this paper need to be adapted for multilingual ecologies including one or more sign languages.

[3] In the remainder of the article we generalize the use of the term "multilingual(ism)" to also include "bilingual(ism)".

dynamics (see, e.g., Lüpke 2016 and Di Carlo et al. 2019). These dynamics are tightly connected to situations in which individuals' multilingual repertoires include mainly "small", highly localized languages, thus leading to the use of the term "small-scale multilingualism".[4]

Crucially for the LD agenda, the "discovery" of small-scale multilingualism has also demonstrated that not all forms of multilingualism are equally implicated in assessing the degree of endangerment of a language. Colonially-mediated dynamics may lead to subtractive multilingualism or other processes of language hierarchization that eventually lead to loss of ancestral languages. By contrast, the indigenous dynamics of pre-colonial origin represent historical continuities, reproducing language ecologies (Haugen 1972) that, in most cases, have in fact favored the maintenance of languages, even of those spoken by relatively small communities (see e.g. Epps 2018 and Cobbinah 2020).

There is another central fact, which Childs et al. (2014) note: "Sociolinguistic contexts are more fragile than lexico-grammatical codes and, therefore, intrinsically more endangered. It is these contexts that will disappear first as smaller communities become transformed by contact with larger ones. Significant lexical data can be collected from even a single 'rememberer'…but documenting a language's sociolinguistic context requires an active speech community" (Childs et al. 2014:172). In other words, small-scale multilingual practices and metapragmatic knowledges must be considered key features that are lost relatively early under endangerment. For this reason, while all forms of multilingualism, even those threatening the survival of endangered languages, are worth being documented, the documentation of small-scale multilingual practices and metapragmatic knowledges seems to us to be all the more urgent.

Conducting documentation of small-scale multilingualism practices can be done only when the documenter targets the lived realities of the speech community, rather than operating from a predetermined notion that their focus should be on the collection of data from a single language (see also Dobrin and Berson 2011). In contrast to the "ancestral-code mode" which is characteristic of most documentary projects, this is what Woodbury refers to as the documentation of the "contemporary communicative ecology" (2011:179). This embodies the idealized "unselective" approach to language documentation, where whatever emerges in daily interactions can potentially be recorded and annotated, irrespective of the languages speakers happen to use (see also Himmelmann 1998:168).

While advocated by some scholars (e.g. Childs et al. 2014), this documentary approach is in fact quite uncommon in LD, as evidenced by an examination of the kinds of projects that are funded by organizations like the Endangered Languages Documentation Programme (www.eldp.net). We believe that it is likely that a major factor in this tendency are uncertainties around how to collect an adequate documentary corpus.

The main goal of this article is to start filling this gap by proposing a foundation for an alternative, viable, and principled view of language documentation that builds multilingual practices into its core. We do this in the hope that more pieces of a shared methodology will follow in the near future—with contributions also from sociolinguists, who have remained somewhat outside of the LD "movement" (see, e.g., Meyerhoff 2019)—and that this will eventually result in more projects focused on such contexts and in a generalized higher concern among scholars and funding agencies for the documentation of small-scale multilingualism phenomena. While our proposals most directly concern the practices of

[4] "Small-scale multilingualism" is, to our knowledge, a term first introduced by Friederike Lüpke (see, e.g., Lüpke 2016). The same kind of phenomena have been referred to differently by other authors, including traditional multilingualism (Di Carlo 2016), endogenous multilingualism (Di Carlo et al. 2019), organic multilingualism (Beyer and Schreiber 2017), and indigenous multilingualism (e.g. Vaughan and Singer 2018). None of these terms are without problems, which is likely why a single term has yet to take hold.

linguistics working on LD, we believe that, if adopted, they would have significant positive impacts in other areas of investigation, such as the study of bilingualism and multilingualism. Moreover, by promoting approaches within LD that put multilingualism at the center of documentation, we also hope that this will help promote expanded dialog between scholars focusing on multilingualism and those working on endangered languages. Finally, within sociolinguistics itself, the present work represents an attempt at providing a principled methodological contribution that we believe is relevant to current developments towards "globalizing sociolinguistics" (see, e.g., the papers in Smakman and Heinrich 2015, especially Meyerhoff and Stanford 2015).

In order to achieve our main goal, we first provide an example taken from a recent detailed study of small-scale multilingualism in an African context (section 2), which will serve as a concrete example of a research situation one may face when documenting small-scale multilingualism. Then, in section 3, we discuss issues of adequacy of documentary corpora, focusing on what appear to have been key obstacles for the development of multilingual LD. In that section, we introduce two keys to this article: the concepts of sociolinguistic adequacy and of indexical space. Section 4 is devoted to outlining the consequences that such a reappraisal has on the treatment of speech data and, in section 5, we summarize the methodology we have applied in our work. The Supplementary Materials provide an attempt to exemplify the methodology through the main lessons learned in a documentary project focused on small-scale multilingualism in rural Cameroon.

2 An example of documented multilingual data

In order to lay out context for the conceptual discussion below, in this section we provide a partly analyzed fragment of a multilingual dialog from the research of Ojong Diba (2019, 2020) in Table 1. Further discussion of the research process through which this data was collected is presented in the Supplementary Materials. The setting for this conversation is the small village of Buu, in the Lower Fungom region of Cameroon (see Good et al. 2011 for a linguistic overview of this region). All the villages mentioned below are located within one hour's walking distance from each other.

The participants are M, T, and V. M (female, ca. 40) is from the village of Buu, T (male, ca. 40) is from the nearby village of Fang and is a hunter, and V (female, ca. 40) is a friend of M's and, like her, comes from Buu. The villages of Buu and Fang are associated with different languages—having ISO 639-3 and Glottolog codes [boe; mund1328][5] and [fak; fang1248], respectively—and are each referred to using the same name as the village. All three participants can communicate in Buu, Fang, and Cameroon Pidgin English (CPE)—which is the lingua franca of the part of Cameroon where Lower Fungom is located—and these are the three languages found in the dialog below. This means that the choice of which language to speak at any given point can be considered socially meaningful.

In the dialog, M and V are chatting in front of M's house in Buu. T stops while passing along with a dead game animal in his hands because M indicates that she would like to buy it. The languages used are: *Fang* (italics), **Buu** (bold), and Cameroonian Pidgin English (*<u>CPE</u>*) (underlined and italics). Close to the turn number, letters identify speaker and addressee, e.g., "M-T" means that "M is speaking to T". Each transcribed turn is provided with a free translation in English, followed by relevant contextual information in some cases (see section 3.3 for an overview of what we mean by "context" here).

[5] In fact, the status of Buu within the so-called "Ji group" (Good et al. 2011), is still unclear, and in other publications on Lower Fungom languages it has been considered a separate language with no current ISO code.

11.	M-T	*ay! ntaya tekem tefebadzie fe me ne tʃi tegwoyagi ŋkamte febə di ifeh... ma nətʃetəgwo*
	Free Tr	No! I will give you 2,200 (francs) ... It is a fair deal.
	Notes	M is negotiating price of the game animal. She uses T's home language, which she learned because of her interactions with Fang speakers (due to the relationships of her father). V can also understand and speak Fang.
12.	T-M	*eme me fie bene, mane tʃiento me yoho yen wum*
	Free Tr	Someone asked to buy this in my village. It is just that I told them that I wanted to go sell it in Wum.
	Notes	T implies that M is asking to buy at an unacceptably low price and that someone offered him more already. Wum is a large nearby town where there would be many more potential buyers.
13.	M-T	*a nte mbagu? an tse tese fele yin abale beli yene?*
	Free Tr	Did you shoot it with a gun? You scattered this part, removed it, and ate with your fufu, didn't you?
	Notes	M is still attempting to buy at a reduced price. She claims that some parts of the animal are missing to justify this.
15.	M-V	*Aunty Mau* **a wule la kpuan?**
	Free Tr	Aunty Mau do you hear the amount he is asking?
	Notes	M turned towards V and is now talking to her directly in Buu, their home language. T can understand Buu.
16.	V-M	**e heh wu**
	Free Tr	I heard
17.	M-V	**ye ntəke fie be kpante kpin la yane bwoeh?**
	Free Tr	He says it is 2,500 francs, is that alright?
18.	M-T	*Sometime edenailuh*?
	Free Tr	Maybe it will be bitter?
	Notes	M turned her face onto T and is now talking to him directly in Fang. The meat of the particular game animal that M would like to buy tastes bitter during a specific part of the year due to the nature of its diet.
21.	M-V	**Bimbe kabed gea fakeh be bentin tsaŋke ndie mise gun bugo tugo ndin neyəŋə** *dey be bringam for me yesterday*
	Free Tr	Won't you give me egusi so that I can cook that cabbage of mine? It was given to me yesterday.
	Notes	M turns back to V speaking in Buu. There is inter-sentential code-switching to CPE in the closing sentence.
22.	V-M	**die be be tsoŋ be noh?**
	Free Tr	cook it with groundnuts, will you?

Table 1: Interaction between M, V, and T (drawn from Ojong Diba 2019:199–205)

Examples like the one above are clearly of documentary interest since they demonstrate the way in which speakers of a set of endangered languages—which is the case for all of the local languages of Lower Fungom (see Supplementary Materials)—deploy multiple languages over the course of their daily lives, which is a critical part of how they are used. A documentary project collecting monolingual data from Lower Fungom would paint a false record of the speech practices of its communities. Moreover, data like this is important for understanding how self-reported information from individuals about their linguistic knowledge and about the way they use multiple languages align with their actual linguistic knowledge and patterns of language use (see, e.g., Mba and Nsen Tem 2020).

This example also demonstrates an important methodological point in terms of the kinds of analyses that are and are not present. There is no interlinear glossing or transcription of tone (even though all the languages make extensive use of tone). Moreover, the segmental transcription is quite rough, and not properly phonemicized.[6] In this regard, the quality of the data falls short of best practice for a "classical" (i.e. monolingual) documentation project.

Nevertheless, important generalizations regarding local linguistic practices can still be derived from Table 1 when it is considered within the entirety of Ojong Diba's corpus. For instance, it includes detailed accounts of the speakers' multilingual repertoires, of their "sociolinguistic life" (including topics such as mobility, special language rights, etc.), and of the space in which the interaction takes place, both from an external and a culture-internal perspective. Therefore, less information in some areas is balanced by more information in other areas.

To be slightly more specific, Example 1 illustrates that Ojong Diba's corpus can allow for the analysis of participants' language choices. While T and V keep using their own "home languages", M carefully selects the language according to whom she is speaking, i.e. Fang with T and Buu with V (CPE *sometime* "perhaps" is best considered a borrowing rather than an instance of code-switching).

This type of interaction exemplifies fundamental traits of Lower Fungom metapragmatic knowledge (see section 3.2 for further discussion of metapragmatics). Throughout the corpus one finds that, regardless of the extent of the multilingual competence of any two speakers, one-to-one interactions are by and large monolingual (cf. Cobbinah et al. 2017 for a very different situation found in a context of small-scale multilingualism in southern Senegal). The choice of the language, when it is anything other than CPE (i.e. when speakers can communicate using a local language), depends on a number of contextual and individual-based factors and cannot be predicted in any general way other than, perhaps, that priority is given to more senior individuals in making the choice. What can be predicted, based on the corpus in Ojong Diba (2019) (see also Ojong Diba 2020:23–25), is that switching between local codes when any two speakers address each other is an atypical choice. It can occur when there is an abrupt change in the immediate context (e.g., as a result of the sudden arrival of a new listener) or to signal specific kinds of social meaning. In particular, it can serve as an attempt to distance one's social connection to the addressee by using a language that signals the lack of a close relationship (see other examples in Ojong Diba 2019, Di Carlo, Good, and Ojong Diba 2019, and Di Carlo, Esene Agwara, and Ojong Diba 2020). The only case in which the semiotic significance of code-switching is somewhat neutralized is when the switch occurs between a local language and CPE. This phenomenon is probably due to the fact that, in the local context, the use of CPE transcends the dense network of possible affiliations with one or the other village-based communities, which are also key for the representation of kinship relations—switching to CPE in other parts of Cameroon, like in francophone areas, might instead convey a radically different, and more significant, social meaning.

[6] Descriptive data about Buu and Fang is relatively limited: see Hombert 1980, Hamm et al. 2002, Good et al. 2011, Ngako Yonga 2013, and Mve et al. 2019 for the major works containing information on these languages that we are aware of.

This example helps clarify the question that lies behind the discussion below: What would LD look like if it began from the assumption that the documentation of multilingualism was at the center of a project rather than at the periphery? In proposing an answer to this question, we should be clear that we do not mean to suggest that projects based on ancestral code approach and that focus on structural linguistic analysis should not move forward since these are also clearly valuable. Rather, we believe that the approach we outline here and the more traditional approach can complement each other and allow us to document languages in a richer way than would be possible by adopting either approach alone.

In the next section, we will try to structure this proposal by discussing and reframing fundamental issues of adequacy with regard to the audience, the scope, and the semiotic order of the contents of LD corpora.

3 Corpus adequacy

3.1 Adequacy and audiences

Summarizing Himmelmann (1998:166), language documentation aims to create "a comprehensive record of the linguistic practices characteristic of a given speech community," that is amenable to further analysis. While language descriptions are generally useful only to grammatically oriented and comparative linguists, suitably annotated documentary corpora have the potential of being of use to a larger group of scholars including, for instance, anthropologists, sociolinguists, discourse analysts, or historians, in addition to linguists.

The issue of corpus adequacy is pertinent for language documentation as a whole (see, e.g., Michael 2011, Woodbury 2011), but it clearly becomes more crucial when the documentary focus includes multilingual speech. The "further analyses" that can be made on such a corpus can potentially derive from one of the many traditions of research on multilingualism in domains as varied as psycholinguistics, language acquisition, the sociology of language, and sociolinguistics. Each have their own levels of analysis, which in its turn are made possible by focusing on certain types of data over others.

The usual intended audience of LD corpora is structurally-oriented linguists rather than sociolinguists, social psychologists of language, psycholinguists, or even speaker communities (cf. Grinevald 2001, Dobrin 2008). It is on the background of their needs, then, that adequacy of an LD corpus has generally been assessed. From the structural linguists' perspective—i.e., where languages are primarily understood as lexico-grammatical codes—a corpus is adequate when it provides data *qua* words, sentences, and texts transcribed phonologically (not phonetically nor in a simplified alphabet) and also analysed grammatically to varying degrees of detail.

The emphasis on adequacy at a lexico-grammatical level can be seen as one major factor accounting for the overall rarity of LD corpora containing multilingual speech data. If an LD project were to target spontaneous linguistic practices, the resulting corpus would include sizable multilingual materials in multiple little-known languages. In order to properly annotate the recordings, linguistic adequacy would require that detailed linguistic knowledge of *all* the languages recorded be accumulated so that phonological transcriptions and morphosyntactic annotations can be produced. However, if gaining enough linguistic knowledge is a demanding task for one undocumented language, it becomes utterly unrealistic when the same is expected for a number of such languages and the time available is bound to the duration of a research project (often coinciding with an individual's doctoral research, see, e.g., Crippen and Robinson 2013).

Facing these kinds of expectations, it is easy to imagine that hardly any linguist would embark on such an enterprise. What structural linguists often fail to appreciate is that there are other options available. Multilingualism can be studied even in the absence of detailed structural understanding of all the languages involved. Scientifically legitimate conclusions for linguistics can be reached, and these are also likely to provide important information and insights on the language-culture nexus (see Michael 2011) that may eventually be beneficial to the linguistic level of analysis itself.

Given this background, before deciding which audience to address in the construction of a corpus and, therefore, what it must possess in order to be adequate, we believe it is first necessary to base these considerations on firm epistemological grounds that are as discipline independent as possible.

3.2 Widening the scope of LD corpora: symbols and indices

There are two main kinds of meaning that signs can convey: indexical and referential (or semantico-referential). Indexical meanings are those that depend on context (e.g. who "I" is, when "now" is, what "that" is, or the meaning of code-switching to a particular language in a particular interaction). Semantico-referential meaning (or function) is referred to with this label because it references "things" and states in the world (making it referential) and because it works based on semantics (i.e. intrinsic, code-dependent meaning) rather than on pragmatics (i.e. context-dependent meaning).

In speech, hardly any sign falls within only one of the above types and produces only one or the other kind of meaning. More commonly, when used in context, linguistic signs are associated with multiple functionalities (see, e.g., Silverstein 1976:45). For instance, there are always multiple ways of conveying the same message, and the choice of how to encode it will inevitably convey some kind of meaning beyond semantico-referential meaning (e.g., whether a request is made via an imperative or an indirect question). In LD corpora, linguists have dealt mostly with semantico-referential meanings—contained in dictionaries, grammars, and texts—and have therefore specialized on the documentation and analysis of language signs *qua* symbols.[7] By contrast, indexical meanings, and the way they are obtained via language signs *qua* indices, do not normally fall within the scope of LD corpora except for those encoded by grammatical items such as demonstratives and other so-called "indexicals" (including pronouns, many circumstantial morphemes, and so on; see Braun 2017 for an overview of indexicals).

This is not an approach that one can adopt when multilingual behaviors are part of an LD corpus, for one simple reason: The use of multiple languages by one and the same speaker, or within a given speech community, is often not required for the production of semantico-referential meaning and is instead connected with the expression of *indexical* meanings primarily related to representations of participants' selves (see, e.g., LePage and Tabouret-Keller 1985 and Irvine and Gal 2000) and of context (see, e.g., the contextualization cues of Gumperz 1982), through the language that they choose to speak. Successful communication at this level is obtained by producing signs from multiple languages whose associated indexical layer of meaning will be decoded by the other interactants in appropriate ways because all of

[7] Following Charles S. Peirce's theory of signs, we can understand signs to be of three main types depending on the relationship existing between the vector (i.e., the signifier) and its meaning (i.e., the signified). Signs in which the vector formally resembles the meaning are called icons: "boom!" and "murmur" are two English examples of icons. Signs whose meaning is connected somehow logically with the vector (i.e., cause-effect, space-time, or context-triggered) are called indices, e.g., deictics like "this" and "that" have meaning only in context or by saying "Yes, we can!" one may evoke Barack Obama's first winning presidential campaign. Signs whose vector is in arbitrary relationship with meaning, mediated through an arbitrary code, are called symbols, e.g., one and the same real item is referred to as "fir (tree)" in English, "sapin" in French, "momi" in Japanese, and "guossa" in Northern Sami.

them share, to varying degrees, a common sense of how to use (signs from) lexico-grammatical codes in culturally appropriate ways. That is, they have a shared metapragmatic knowledge—not just a shared grammar—and at least some portions of "indexical space" in common (see section 3.3 below)—not just a common lexicon. In many small communities, a key feature of metapragmatic knowledge is when the use of a given language is called for.

We begin then with an assumption that a key goal in collecting an LD corpus reflecting the multilingual behaviors of a community is to *collect the data that is required to understand the indexical value of language choice in interaction*. If this is the case, what, then, should an LD corpus of multilingual data contain? We explore this question in the next section.

3.3 Contents of LD corpora: A re-appraisal and the notion of "indexical space"

A documentary corpus typically includes speech recordings (in the form of both video and audio files), annotations on the recordings (such as transcriptions, translations, etc), metadata at the level of the recording session (e.g. date and place of recording, participants recorded, etc.), and a collection of analytical statements that, taken together, help users make sense of what is referenced in the annotations (typical examples are dictionaries and sketch grammars). Theoretically, there are no limits as to the domains of knowledge that are covered in annotations and analytical statements, which can include observations about ethnographic, geographic, sociolinguistic, musicological, or even ethological aspects of interest found in the recordings. To date, however, LD corpora contain mostly speech data with linguistic annotations only.

Due to the fact that multilingual behaviors foreground the indexical dimension of meaning-making of linguistic signs (as discussed just above in section 3.2), corpora containing recordings of multilingual speech will have to also provide annotations and analytical statements that allow a user to make sense of how indexical meanings are produced. Dictionaries and grammars enrich the corpus user by illuminating the relationships existing between, on the one hand, linguistic signs and the "universe of the nameable" (i.e. lexicon) and, on the other, between linguistic signs themselves (i.e. grammar). Likewise, when the linguistic signs recorded instantiate indexical relationships with elements found in the context of the speech event, then *an adequate corpus is one that includes not only speech data but also sufficient reliable information about the contexts and items that are indexed*. It is important to note here that we speak of "contexts" in the plural. Inspired by Goodwin and Duranti (1992), we identify two main types of context which documentary efforts should be directed to: situational and extra-situational. In rough terms, the former refers to the locale in which a recorded interaction takes place. The latter includes both metalinguistic and metapragmatic knowledges—and therefore a sizable amount of ethnographic knowledge about the speech community—as these are highly influential in determining how people shape their multilingual behaviors. For this reason, data at both levels should be collected and included in the documentary corpus. The universe of what can be indexed linguistically using a certain kind of metapragmatic knowledge is what we call here its "indexical space" (see Di Carlo forthc. for more details on the notion of indexical space).[8]

Issues concerning the adequacy of corpora with regard to how much of the indexical space they must capture and the tools to be used will not be considered in detail here (though see

[8] The concept of indexical space is clearly related with Eckert's "indexical field" (Eckert 2008) though it may be conceived of as being the "container" of all the possible indexical fields that can be activated in a given language ideology. Otherwise stated, indexical space is here intended as *the universe of whatever can be indexed through linguistic means*, and therefore as the domain of human cognition in which processes of indexical ordering take place (Silverstein 2003).

section 5.3 for an overview). What is more important in the present context is the observation that, for an LD corpus of multilingual data to be adequate, annotations and analytical statements cannot be limited only to the linguistic level but should also cover a broad range of contextual factors (e.g. ethnographic, sociolinguistic, etc. following the seminal proposals of Hymes (1972[1986]) for an ethnography of speaking, including its SPEAKING mnemonic, which lies at the foundations of the approach presented here; see also see also contributions in this volume by Lüpke, Sagna & Hantgan, and Walworth et al. as instances of studies aimed at understanding the diversity of multilingual practices in small-scale multilingual settings through multidisciplinary methods, with an emphasis on ethnography). Depending on the levels of annotation and analyses that are needed, the ways in which multilingual speech data are to be collected and approached will need to be adjusted. We develop this point further in section 4 and section 5 (see also Supplementary Materials).

4 Multilingual speech data

Multilingual speech in natural conversation is manifested in phenomena such as borrowing, calquing, and code-switching. Drawing on some of the methods commonly used to analyse these phenomena, we can identify linguistic, sociolinguistic, and pragmatic levels of analysis. Linguistic analyses can be focused on lexical choices, or on the syntax, morphology, and phonology of multilingual speech. Sociolinguistic and pragmatic levels of analysis can be focused on language attitudes, social indexicality, language vitality, style, turn-taking, phenomena connected to accommodation, and metapragmatic knowledge.

However different from each other they are as to the type of linguistic knowledge that they build upon, at a closer inspection one realizes that knowledge of the indexical space is necessary at all of these levels due to the fact that multilingual speech is inherently indexical in nature, as discussed in section 3. Different degrees of need for detailed linguistic knowledge across levels of analysis are connected mainly to matters of the scale of the analytical units. When the observable vector of an index is "a language as a whole"—as it is, for example, in many instances of code-switching phenomena—then an initial analysis does not *need* to be based on the observation of micro-phenomena—such as phonological variants—but can be performed on entire chunks of speech flow, from sentences to speech turns to even entire speech events, as was done in the example presented in section 2.

Of course, providing a phonological transcription would be of great value for supporting more in-depth analyses. However, we want to emphasize that the enormous efforts required to obtain it would radically jeopardize the *documentation* project whose success, in fact, does not depend on such linguistic detail but, rather, on details about the indexical meanings produced during the interaction, and this can be achieved even without a phonological transcription. A key claim we are making here, therefore, is that, in LD corpora of contemporary communicative practices in contexts of small-scale multilingualism, a general understanding of the conceptual space that is being used by speakers to produce indexical meanings can be reached *before* embarking on the task of obtaining detailed grammatical knowledge. Within this domain, priority must be given to matters such as the identification of the codes used, the repertoires available to interactants, and the social significance of the codes present in the indexical space before substantial efforts are made to increase the grammatical knowledge of the codes involved (if any such efforts can be made).

Such requirements, taken together, comprise what we refer to, for the sake of convenience, as "sociolinguistic adequacy". The documentary model that we develop here is intended to ensure that LD corpora of multilingual speech can meet such a threshold.[9] Structural linguistic adequacy in such cases is, we believe, less urgent because, without knowledge of the indexical space, multilingual behaviors will hardly be interpretable and, due to the lack of reliable annotations about the indexical meanings produced in speech data, the corpus will be less amenable to further analyses. In addition to being less urgent, achieving structural linguistic adequacy in contexts of small-scale multilingualism is often an unrealistic goal.

These things being said, we do not intend to suggest that sociolinguistic and structural linguistic adequacy should be taken as two opposite, mutually exclusive approaches. Rather, we view them as two cycles of a larger process. This recalls the words of Grinevald (2001:288) about the interplay between descriptive and documentary approaches in language documentation. She concludes that "the process would start with an initial description, this description becoming essential for a wider type of documentation, which itself will allow for a more sophisticated and comprehensive description, and so on." While she focuses on the relationship between documentation and description, we are, instead, emphasizing two different approaches to documentation itself. For multilingual documentation, the key phase is the one in which sociolinguistic adequacy, instead of structural linguistic adequacy, is at the forefront: This will then provide essential background for a more detailed description of multilingual speech data, followed by more in-depth sociolinguistic analyses, which will feed improved linguistic analyses, and so on.

5 Achieving sociolinguistic adequacy

5.1 The challenge of "relinquished control"

One key underpinning of sociolinguistic adequacy, especially when projects are conducted in contexts of small-scale multilingualism, is that it entails a greater reliance on speaker consultants for the production of primary data (i.e. minimally processed data such as free translations or identification of the codes used in the recordings) rather than of raw data only (i.e. speech).[10] Standard approaches to structural linguistic analysis may rely on the analyses of raw data produced by consultants, for instance in the form of careful re-pronunciations of naturalistic speech to assist in transcription or free translations produced by them. However, these are generally "filtered" to create primary linguistic data through long-standing analytic techniques such as those that support phonemic and morphological analysis.

By contrast, such filtering would be inappropriate for a corpus aiming for sociolinguistic adequacy where the analyses that speakers produce are not merely a tool for the linguist to arrive at the "right" structural analysis but, rather, are, in and of themselves, a key kind of data for understanding locally salient indexical meanings. In practice, this "unfiltering" is most readily observed in instances where consultants are asked to comment on recordings of their (multilingual) linguistic practices. In addition to identifying other participants in the recording and the named varieties used in the interaction, the consultant can also be asked about the motivations that led them to use or switch to a certain language. Of course, not all information collected in this way should be taken at face value, since folk rationalizations may or may not be revealing of metapragmatic knowledge, which is the ultimate target of sociolinguistic adequacy. Nonetheless, language choice in interaction is often subject to conscious metapragmatic manipulation and, for this reason, speakers may not only be aware of their motivations, but also able to articulate them, thereby producing primary data that the

[9] We must immediately acknowledge that this is a practically and theoretically complicated notion that is certainly in need of more detailed exploration than can be provided here.

[10] We follow Himmelmann's (2012) definitions of raw and primary data in language documentation.

researcher will not filter through an existing “grammar” but, rather, integrate into the corpus in various ways—e.g., via session-based or topic-based notes, annotations to the recording, etc.[11]

In contexts of small-scale multilingualism, common low-control research activities are those in which the researcher not only needs to rely on the help of native speakers to make recordings (raw data) and produce annotations (primary data), but also cannot directly check their accuracy (i.e., no filtering is possible). This could occur, for instance, when a researcher must rely on a consultant’s judgment that a given stretch of discourse involves the use of a number of different “languages” which the researcher has not had much exposure to, as was the case, for instance, in the analysis of the data presented in Table 1 in section 3. The evident shortcoming of such a situation of “relinquished control”, well described by Grinevald (2001: 301) as “one of the most difficult constraints for academics to accept” is the risk of not being able to guarantee data validity (see, e.g., Tillery 2000). However, this potential weakness can be countered in two ways: (i) by having multiple consultants do the same kind of work for the same data and then comparing the results for misalignments between them, which can then be addressed in various ways, such as focus group sessions, and (ii) by sampling local metapragmatic knowledge indirectly through tools such as the Matched-Guise Test technique (see Chenemo and Neba 2020 for an interesting adaptation of this tool to a rural African context).

5.2 The need for an ethnographic approach

A fundamental consideration to make at this point concerns what kinds of research methods can still be considered “high-control activities” in such a research scenario. For one thing, these are key for the scholarly world to accept the model proposed here. The kind of documentation we are proposing does not mean recording whatever people say and collecting whatever accounts they offer of their behaviors indiscriminately. For sociolinguistic adequacy to be met, it is essential that the approach is ethnographic in kind, which means researchers should be trained in how to identify and understand as much as possible about the speakers’ own perspectives.

This becomes evident, for instance, in the identification of the codes used by speakers, an activity which should be rooted in the concept of local saliency prior to any possible categorization facilitated by the use of linguistic comparative tools—whose aim is to assess the distance between codes and, therefore, “establish” whether these are, e.g., “separate languages” or “just dialects of a single language”. If a code is named and is systematically identified in recordings by consultants, then it must be documented as an element of the community’s linguistic repertoire, regardless of its degree of similarity with any other code present in the repertoire (see also Esene Agwara 2020:189 and Khanina, this volume, for a similar ethnographic exercise carried out in Siberia).[12]

Another example of what an effective ethnographic approach may entail in practice concerns the speakers themselves. These, too, must be “documented” as their available roles and identities are likely to shape the way they speak with other community members. The basic (auto-)ethnographic recognition here is that salient features of personal and social identity are culture specific and, therefore, initially inaccessible to outside researchers. For

[11] While not focused on multilingualism specifically, Silverstein 2001[1977] remains a key reading with regard to fieldwork of the kind we advocate here.

[12] This does not necessarily equate with saying that all codes will need to have a separate indexical value, as this assumption would carry a risk that some situations would be represented in an infelicitous way (see, e.g., Meeuwis and Blommaert 1998).

this reason, speakers should not be categorized solely according to widely-used sociological parameters such as age, gender, socio-economic status, and occupation. Rather, additional efforts are required in order to collect information that will initially be particularly wide ranging and encompass topics that do not immediately lead to systematization—such as people's life histories and kinship and other relations within and outside of the community—and which can be narrowed down to a set of more salient features—such as the provenance of people's personal names (see, e.g., Di Carlo and Good 2014:249–250, Esene Agwara 2020:189–190, and Di Carlo forthc.)—only when more data becomes available.

What we have said in this section so far brings about one first (i.e., not the only) requirement for those embarking in the creation of a multilingual LD corpus: Anyone "preparing for fieldwork should read the contemporary ethnographic literature on the broader region in which they plan to work" (Dobrin 2008: 317). Not doing so would risk failing to see many important areas of data collection that have already been identified by previous work.[13]

5.3 An overview of the approach proposed

Based on the discussion above, we provide an overview of the data collection approach that we advocate in Table 2, including possible associated research outputs and activities. The abbreviations *H*, *L*, and *HL* refer to the degree of control researchers have on the corresponding activity: H = High control, L = Low control, HL = High control on tools, Low control on output (i.e. strategies are needed to maximize data validity, see section 5.1 above).[14]

Domain	Target	Output	Activity	Control
Indexical space	Speakers	Individual-based sociolinguistic profiles (speaker metadata) References to profiles in annotations on recordings	Structured interviews covering topics such as an individual's multilingual repertoire, history of mobility, social networks, and social status according to local norms	HL
	Situational contexts	Ethnographic notes describing speech events: setting, participants, norms, ends, genres, etc. References to the notes in annotations on recordings	Observation of and note-taking about speech events both directly and during collaborative work sessions with speakers Description that is both etic (i.e. external, "objective") and emic (i.e. culture-internal)	HL
	Extra-situational contexts (including metapragmatic knowledge)	Ethnographic notes about topics such as language ideologies, ethnohistories, spiritual beliefs, settlement patterns, social and	Interviews and focus groups on cultural features: These may be very general (e.g. ethnohistory, spiritual beliefs) or specific to the speech events recorded (e.g.	HL

[13] Relatedly, we would also like to note that adopting the specific ethnographic method of participant observation, at least during part of a research period, is likely to yield valuable insights that cannot be easily obtained by using traditional linguistic methods alone. See Dobrin and Schwartz (2016:260–264) for relevant discussion in a documentary context. See also Morozova & Rusakov, Khachaturyan & Konoshenko, and Vydrina, this volume, for examples of participant observation used in studies of multilingualism.

[14] In many documentary projects, speakers themselves are in fact often the ones making recordings using equipment provided by the researcher. In such cases, this activity would be highly controlled by the researcher at the level of choosing the equipment, but not with respect to how it us used on the ground. We indicate this using the categorization "H (HL)" for the activity of recording in Table 2.

		physical spaces References to the notes in annotations on recordings	myths, social etiquette, etc.) Design, administration, and analysis of tools for the exploration of language attitudes and ideologies, like the Matched-Guise Technique (Lambert et al. 1960; see Chenemo and Neba 2020 for an adaptation of the same technique to contexts of small-scale multilingualism)	
Speech	Recordings	High-quality audio and video recordings	Choice and use of equipment during recording sessions	H (HL)
		Comprehensive speaker and session metadata	Choice of metadata elements and encoding schemes	H
	Speech data	Annotations on recordings allowing identification of codes used by speakers (at different levels of detail)	Collaborative work sessions with speakers	L
		Accessible transcription of speech without emphasis on phonemicization or standardization	Choice of approach to transcription and system to use Collaborative work sessions with speakers	HL

Table 2: Research strategies for achieving sociolinguistic adequacy across domains, targets, outputs, activities, and degree of researcher control

6 Conclusion

In this article, we have tried to re-imagine LD from the perspective of those who plan to target contexts of small-scale multilingualism. The key step in this process has been an appraisal of the larger-than-usual weight that must be given to the indexical function of linguistic signs in a corpus containing multilingual speech data, as compared to more typical monolingual LD corpora. This has led us to propose that such a corpus will have to include both speech data (i.e. recordings of speech events) as well as information about what we called here the "indexical space"—i.e. speakers' metadata plus situational and extra-situational context, which together amount to the ideal universe of whatever can be indexed using language through the metapragmatic knowledge that is shared by members of a given speech community. Our claim has been that "sociolinguistic adequacy" must be prioritized in the creation of multilingual LD corpora over structural linguistic adequacy, and we have summarized its main theoretical (section 5) and practical aspects (see Supplementary Materials).

We want to emphasize here that a sociolinguistically adequate documentary corpus, such as this one, makes it possible to gain access to information that can feed many other types of analysis, not only sociolinguistic, but also linguistic, anthropological, and historical. This can be seen, for example, in the insights on strategies of identity construction through language choice that are discussed in Di Carlo, Good, Ojong Diba (2019:§5) or the historical and ethnographic interpretation of small-scale multilingualism in the Casamance region of Senegal found in Lüpke (2018). Crucially, it could also inform efforts at revitalization as the local language ecology that is being documented is clearly connected with the maintenance of small languages.

A final remark should be made with respect to the intended audience of this article. As with most methodological proposals, those whose work would be most directly impacted is a relatively small set of specialists, in this case scholars whose research focuses on the documentation of endangered languages. At the same time, the research situations discussed above—i.e., contexts of small-scale multilingualism—will be of interest to a much wider range of scholars, such as sociolinguists who, by and large, are not normally involved in the creation of LD corpora. We hope others, especially scholars whose research focuses on multilingualism, even if it has been primarily oriented towards "large-scale" societies, will continue a conversation we have only started here, fill gaps we have left open, and point out ways in which the approach outlined above needs to be refined. We also hope that our proposal will have the consequence of reaffirming the *need* for a multidisciplinary and team-based approach to language documentation (*pace* Austin and Grenoble 2007:22), especially when the documentary task is as complex as the one we have focused on in this article.

References

Austin, P. K. and L. A. Grenoble. 2007. Current trends in language documentation. In P. K. Austin (ed.) *Language documentation and description*, volume 4, 12–25. London: Hans Rausing Endangered Languages Project.

Auer, P. 1999. From code switching via language mixing to fused lects: Towards a dynamic typology of bilingual speech. *International Journal of bilingualism* 3(4): 115–158.

Beyer, K. and H. Schreiber. 2017. Social network approach in African sociolinguistics. *Oxford Research Encyclopedia of Linguistics*. https://dx.doi.org/10.1093/acrefore/9780199384655.013.236

Braun, D. 2017. Indexicals. In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/sum2017/entries/indexicals/.

Chenemo, M. N. and A. N. Neba. 2020. Essentialism and indexicality in a multilingual rural community: The case of Lower Bafut in North-West Cameroon. In P. Di Carlo and J. Good (eds.) *African Multilingualisms*, 229–259. Lanham, MD: Lexington Books.

Childs, T., J. Good, and A. 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, A., A. Hantgan, F. Lüpke, and R. Watson. 2017. Carrefour des langues, carrefour des paradigmes. In M. Auzanneau, M. Bento, and M. Leclère (eds) *Espaces, mobilités et éducation plurilingues: Éclairages d'Afrique ou d'ailleurs*, 79–97. Paris: Édition des Archives Contemporaines.

Cobbinah, A. 2020. An ecological approach to ethnic identity and language dynamics in a multilingual area (Lower Casamance, Senegal). In P. Di Carlo and J. Good (eds.) *African Multilingualisms. Rural linguistic and cultural diversity,* 71–105. Lanham, MD: Lexington Books.

Connell, B. 2009. Language diversity and language choice: A view from a Cameroon market. *Anthropological Linguistics*, 51, 2: 130–150.

Crippen, J. A. and L. C. Robinson. 2013. In defense of the lone wolf: Collaboration in language documentation. *Language Documentation & Conservation* 7: 123–135. http://hdl.handle.net/10125/4577.

Di Carlo, P. 2011. Lower Fungom linguistic diversity and its historical development: Proposals from a multidisciplinary perspective. *Africana Linguistica* 17: 53–100.

Di Carlo, P. 2016. Multilingualism, affiliation and spiritual insecurity. From phenomena to processes in language documentation. In M. Seyfeddinipur (ed.). *African language* *documentation new data, methods and approaches. Language Documentation & Conservation* special publication no. 10: 71–104. http://hdl.handle.net/10125/24649.

Di Carlo, P. 2018. Towards an understanding of African endogenous multilingualism: Ethnography, language ideologies, and the supernatural. *International Journal of the Sociology of Language* 254: 139–163.

Di Carlo, P. forthcoming. Reappraising questionnaires in the study of multilingualism: lessons from contexts of small-scale multilingualism. To appear in L. Grenoble and J. Martin (eds.) *Multilingualism, contact, and documenting endangered languages*, special volume of the *Journal of Language Contact*.

Di Carlo, P. and Good, J. 2014. What are we trying to preserve? Diversity, change, and ideology at the edge of the Cameroonian Grassfields. In P. Austin and J. Sallabank (eds.). *Endangered languages: Beliefs and ideologies in language documentation*, 229–262. Oxford: Oxford University Press.

Di Carlo, P. and J. Good (eds.). 2020a. *African multilingualisms*: *Rural linguistic and cultural diversity*. Lanham, MD: Lexington Books.

Di Carlo, P. and J. Good. 2020b. Introduction: Understanding the diversity of multilingualisms in Sub-Saharan Africa. In P. Di Carlo and J. Good (eds.) *African multilingualisms*: *Rural linguistic and cultural diversity*, xv–xxxvii. Lanham, MD: Lexington Books.

Di Carlo, P., J. Good, and R. Ojong Diba. 2019. Multilingualism in rural Africa. *Oxford Research Encyclopedia of Linguistics*. https://dx.doi.org/10.1093/acrefore/9780199384655.013.227.

Di Carlo, P., A. Esene Agwara, and R. Ojong Diba. to appear. Multilingualism and the heteroglossia of ideologies in Lower Fungom (Cameroon). Accepted for publication in *Sociolinguistic Studies*, special volume entitled *Urbanized African sociolinguistics–Questioning research foci.*

Dobrin, L. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment from Melanesia. *Language* 84: 300–324.

Dobrin, L. and J. Berson. 2011. Speakers and language documentation. In P. Austin and J. Sallabank (eds.) *The Cambridge handbook of endangered languages*, 187–211. Cambridge: Cambridge University Press.

Dobrin L. and J. Schwartz. 2016. Collaboration or participant observation? Rethinking models of 'linguistic social work'. *Language Documentation & Conservation* 10: 253–277. http://hdl.handle.net/10125/24694.

Eckert, P. (2008). Variation and the indexical field. Journal of SocioLinguistics, 12(4), 453–476. https://doi.org/10. 1111/j.1467-9841.2008.00374.x

Epps, P. 2018. Contrasting linguistic ecologies: Indigenous and colonially mediated language contact in northwest Amazonia. *Language and Communication* 62, part B: 156–169.

Esene Agwara, A. 2013. *Rural multilingualism in the North West region of Cameroon*. Buea, Cameroon: University of Buea MA thesis.

Esene Agwara, A. 2020. What an ethnographically informed questionnaire can contribute to the understanding of traditional multilingualism research: Lessons from Lower Fungom. In P. Di Carlo and J. Good (eds.) *African multilingualisms: Rural linguistic and cultural diversity,* 183–206. Lanham, MD: Lexington Books.

Evans, N. 2008. Review of Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.) Essentials of Language Documentation. *Language Documentation & Conservation* 2: 340–350.

Good, J., J. Lovegren, J.P. Mve, C. Nganguep, R. Voll, and P. Di Carlo. 2011. The languages of Lower Fungom region of Cameroon: Grammatical overview. *Africana Linguistica* 17: 101–164.

Goodwin, C. and A. Duranti. 1992. Rethinking context: An introduction. In A. Duranti and C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon*, 1–42. Cambridge: Cambridge University Press.

Grinevald, C. 2001. Encounters at the brink: Linguistic fieldwork among speakers of endangered languages. In O. Sakiyama (ed.), *Lectures on endangered languages: 2*, 285–313. Kyoto: Nakanishi Printing Company.

Hale, K., M. Krauss, L. J. Watahomigie, A. Y. Yamamoto, C. Craig, L. M. Jeanne, and N. C. England. 1992. Endangered languages. *Language* 68(1): 1–42.

Hamm, C., J. Diller, K. Jordan-Diller, and F. Assako a Tiati. 2002. A rapid appraisal survey of Western Beboid languages (Menchum Division, Northwest Province). SIL Electronic Survey Reports 2002-014. Yaoundé: SIL International.

Himmelmann, N. P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195.

Himmelmann, N. P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation* 6: 187-207.

Hombert, J.-M. 1980. Noun classes of the Beboid languages. In Larry M. Hyman (ed.), *Noun classes in the Grassfields Bantu borderland*, 83–98. Southern California Occasional Papers in Linguistics 8. Los Angeles: University of Southern California Department of Linguistics.

Hymes, D. 1972[1986]. Models of the interaction of language and social life. In J. J. Gumperz and D. Hymes (eds.), *Directions in sociolinguistics: The ethnography of communication*, 35–71. Oxford: Basil Blackwell.

Irvine, J. T. and S. Gal. 2000. Language ideology and linguistic differentiation. In P. V. Kroskrity (Ed.), *Regimes of language: Ideologies, polities, and identities*, 35–84. Santa Fe, NM: School of American Research Press.

Lambert, W. E., R. Hodgson, R. C. Gardner, and S. Fillenbaum. 1960. Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology* 60(1): 44–51.

Le Page, R. B. and A. Tabouret-Keller. 1985. *Acts of identity: Creole-based approaches to language and ethnicity*. Cambridge: Cambridge University Press.

Lovegren, J. 2013. *Mungbam grammar*. Buffalo, NY: University at Buffalo PhD dissertation. http://hdl.handle.net/10477/50597.

Lüpke, F. 2016. Uncovering small-scale multilingualism. *Critical Multilingualism Studies*, 4(2): 35–74.

Lüpke, F. 2018. Multiple choice: Language use and cultural practice in rural Casamance between convergence and divergence. In J. Knörr and W. T. Filho (eds.) *Creolization and pidginization in contexts of postcolonial diversity*, 181–208. Leiden: Brill.

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

Meeuwis, M. and J. Blommaert. 1998. A monolectal view of code-switching: Layered code-switching among Zairians in Belgium. In P. Auer (ed.) *Code-switching in Conversation: Language, Interaction, and Identity*, 76–100. London: Routledge.

Menang, T. 2004. Cameroon Pidgin English (Kamtok): Phonology. In E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton (eds.), *Handbook of varieties of English: Volume I: Phonology*, 902–917. Berlin: Mouton.

Meyerhoff, M. 2019. Unnatural bedfellows? The sociolinguistic analysis of variation and language documentation. *Journal of the Royal Society of New Zealand* 49(2): 229–241.

Meyerhoff, M. and J. N. Stanford. 2015. Tings change, all tings change: The changing face of sociolinguistics with a global perspective. In D. Smakman and P. Heinrich (eds.), *Globalising sociolinguistics: Challenging and expanding theory*, 1–15. London: Routledge.

Michael, L. 2011. Language and Culture. In In P. Austin and J. Sallabank (eds.) *The Cambridge handbook of endangered languages*, 120–140. Cambridge: Cambridge University Press.

Mve, P. 2013. Aspects of the phonology of Fáŋ. Yaoundé, Cameroon: University of Yaoundé MA thesis.

Mve, P. J., N. C. Tschonghongei, P. Di Carlo, and J. Good. Cultural distinctiveness and linguistic esoterogeny: The case of the Fang language of Lower Fungom, Cameroon. In P. W. Akumbu and E. P. Chie (ed.), *Engagement with Africa: Linguistic essays in honor of Ngessimo M. Mutaka*, 163–178. Köln: Köppe.

Myers-Scotton, C. 1993. *Social motivations for code switching: Evidence from Africa*. Oxford: Clarendon Press.

Ngako Yonga, M. D. 2013. Ébauche phonologique et morphologique de la langue bu. Yaoundé, Cameroon: University of Yaoundé MA thesis.

Ojong Diba, R. 2019. The sociolinguistic dynamics of rural multilingualism in Africa: The case of Lower Fungom. Buea, Cameroon: University of Buea PhD thesis.

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

Rumsey, A. 2018. The sociocultural dynamics of indigenous multilingualism in northwestern Australia. *Language and Communication* 62, part B: 91–101.

Silverstein, M. 2001. The limits of awareness. In A. Duranti (ed.) *Linguistic anthropology: A reader*, 382–401. Malden: Blackwell.

Silverstein, M. 2003. Indexical order and the dialectics of sociolinguistic life. *Language and Communication* 23: 193–229.

Smakman, D. and P. Heinrich (eds.). 2015. *Globalising sociolinguistics: Challenging and expanding theory*. London: Routledge.

Tillery, J. 2000. The reliability and validity of linguistic self-reports. *Southern Journal of Linguistics* 24: 55–68.

Vaughan, J. and R. Singer. 2018. Indigenous multilingualisms past and present. *Language and Communication* 62, part B: 83–90.

Voll, R. 2017. A grammar of Mundabli A Bantoid (Yemne-Kimbi) language of Cameroon Leiden: Leiden University PhD thesis.

Woodbury, A. C. 2005. Ancestral languages and (imagined) creolisation. In P. K. Austin (ed.), *Language documentation and description*, *volume 3*, 252–262. London: SOAS.

Woodbury, A. C. 2011. Language documentation. In P. K. Austin and J. Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.

Towards a coherent methodology for the documentation of small-scale multilingualism: Dealing with speech data

SUPPLEMENTARY MATERIALS

Pierpaolo Di Carlo
University at Buffalo

Rachel A. Ojong Diba
University of Buea

Jeff Good
University at Buffalo

From recommendations to practice

In this supplementary document, we aim to provide interested readers with more information about the research process we have followed in our work. In order to do this, we first return to the example of annotated multilingual speech presented in section 2 of the article and discuss the process through which the second author collected and analyzed it (section A). In section B, we provide an example of what we mean by "speaker's metadata" (see Table 2 and section 6 of the article). In section C, we provide the ethnographically-informed interview guide that we used in order to collect "speaker's metadata". Our goal is, on the one hand, to make clear the ways in which the recommendations contained in section 5 of the article emanate from concrete experience in the field, and, on other hand, to help illustrate the ways in which adopting these recommendations can be done in a manageable way.

A.1 Target area and themes

The data in section 2 is drawn from Ojong Diba (2019), which discusses in detail the overall research project that resulted in its collection, and we discuss the approach adopted in that work in this section in order to clarify the kinds of steps that can be taken to produce sociolinguistically adequate documentation. In this particular case, the target region for the work, Lower Fungom, had been the subject of a number of previous surveys covering the linguistic (Hombert 1980, Hamm et al. 2002, Good et al. 2011) and ethnographic (Di Carlo 2011, Di Carlo and Good 2014) situation.

From these previous linguistic and ethnographic studies, Ojong Diba (2019) was able to draw valuable insights for designing a project to collect and analyze multilingual data. Key features that were already known include:

- The core of the Lower Fungom linguistic area is about the size of the city of Amsterdam, and about 10,000 people live in thirteen villages traditionally considered independent chiefdoms.
- With seven different non-Bantu Bantoid languages associated with the region, it is an area of very high language diversity.
- Local language ideologies tend to stress a one-to-one relationship between villages and languages. While a linguist would identify seven languages (five single-village languages and two language clusters), a local resident would recognize thirteen different "talks", each going by the name of the village it is associated with, though similarities among some of them (what a linguist would call language clusters) would also be readily acknowledged locally.

On the sociolinguistic side, a survey of around 100 individuals was carried out in 2012, mostly involving self-reports about individual multilingualism via an ethnographically-informed, semi-structured interview guide (Esene Agwara 2013). The main results were as follows (where the term *lect* is used for any named linguistic variety irrespective of its classification as a language or dialect—see Di Carlo et al. to appear: sec. 3.5):

- None of the respondents (n = 97; male 58%, female 42%; ages 16–80) were monolingual; the minimal repertoire included one local language and Cameroon Pidgin English.
- On average, men reported being able to understand about ten lects and to speak six of these, while women reported being able to understand eight lects and to also be able to speak six of these.
- Multilingual repertoires seem to be more extensive (reportedly up to seventeen lects, including lects associated with Lower Fungom villages as well as villages outside the region) in older generations; if confirmed, this might be seen as a consequence of the growing role of Cameroon Pidgin English in daily life since the 1960s.
- Language choice in the local lects seems determined by (multiple) village and social network affiliations rather than by essentialist identities linking language to a notion like "ethnicity" (see Di Carlo and Good 2014).

This self-reported information provided valuable background for the research on language use reported in Ojong Diba (2019). The goal of this research was to document contemporary linguistic practices in order to understand how select speakers of Lower Fungom languages deploy their multilingual repertoires in interaction.

We realize that this amount of information may not be available in most contexts where LD projects are carried out, and this should not be taken as a prerequisite. For one thing, one can still aim for sociolinguistic adequacy even if the research area is poorly studied from a sociolinguistic standpoint. The main difference would be the scope of the analyses that the corpus would support. Furthermore, results that have been obtained for one area—such as those summarized in, e.g., Chenemo and Neba 2020, Esene Agwara 2020, and Ojong Diba 2020 for Lower Fungom and nearby areas—can be useful as guides for research in other nearby areas. Finally, it must be kept in mind that significant ethnographic knowledge on a given speech community or its broader region may be found in the anthropological, rather than the linguistic, literature. It is also clear that the approach we advocate here can be most effectively achieved via multidisciplinary collaboration.

That being said, we would recommend that any study of multilingual practices include a sociolinguistic survey component so that at least self-reported information can be gathered to inform other aspects of data collection. Even a small survey of, for instance, ten to twenty individuals is likely to lead to usable results, provided it is broad enough in scope to extend ethnographic knowledge of the speech community (see section 5.2 above; potentials for complementarity in documentary agendas are further discussed in Childs et al. (2014) and Di Carlo (forthc.)).

A.2 Research activities

A.2.1 Corpus theorization

The following concerns were considered by Ojong Diba (2019) when developing the corpus discussed above. That is, they were the foundation of a kind of corpus theorization (see Woodbury 2011: 180–183).

1. Since African multilingualism is known mostly from urban contexts, the corpus had to provide a sample of multilingual practices (in the form of audio and video recordings) that were meaningful for its specific rural context, in order to ensure it had the highest possible documentary value.

2. A properly representative sample required capturing ethnographically important information. As a result, the corpus had to include not only recordings of an informed selection of speech events, but also ethnographic information in the form of self-reported information, data from interviews, and direct observations.
3. Availability of ethnographic information in the corpus would make it further possible to explore local metapragmatics and language ideologies. These analyses would contribute to the understanding of the local indexical space, thus benefitting possible future analyses.
4. Numerous languages are used by multilinguals in Lower Fungom, realistically more than ten local lects, only few of which have been researched in detail (see, e.g., Lovegren 2013, Voll 2017), in addition to English, Cameroon Pidgin English and, possibly, some French, as well as other local languages associated with areas near Lower Fungom. As a consequence, the kind of linguistic knowledge that would be necessary to provide the corpus with detailed structural linguistic annotations—such as phonological transcription, morphological glosses, a lexicon, etc.—is unachievable within the scope of dissertation research. The transcription method had to be selected in response to sociolinguistic adequacy requirements rather than structural linguistic adequacy, and data would have to be analyzed by "relinquishing control", i.e. relying extensively on local consultants without being able to rely on long-established techniques for "filtering" this kind of data, as is standard for fieldwork aimed at structural linguistic analyses (see also section 5.1).
5. Relinquishing control raises the risk that the collected data cannot be considered fully valid for research. In order to ensure the validity of the data, collaboration with multiple local consultants was required, with them working on the same recordings as a way to verify the accuracy of their analyses.
6. The annotations on the corpus should make the most out of the availability of ethnographic and self-reported information. This requires a way to allow complex annotations to be made available to corpus users in an effective way.
7. Given the different types of data included, the corpus should be associated with a transparent and rich metadata apparatus distinguishing between no less than two sets of data, namely speech data (i.e., metadata for the recordings) and data relevant to understanding the indexical space (e.g., metadata describing speakers). While the former could be encoded using existing models, the latter is likely to require setting-specific adaptations.

A.2.2 Corpus design and creation

Having established a number of parameters that needed to be considered in the design of the corpus, corpus creation activities conducted by Ojong Diba (2019) concretely took place through the following steps.

Collection of speaker metadata and selection of consultants
Before starting to make recordings, a feasibility study was conducted based on interviews of more than thirty people from different villages of Lower Fungom using an ethnography-based sociolinguistic interview guide (developed on the basis of what was used in Esene Agwara 2013, see Di Carlo forth. for more details) and three primary multilingual consultants were chosen to be the focus for data collection. These individuals, all from the area and born in different villages, were not only representative of highly multilingual individuals (speaking at least nine lects including English and Cameroon Pidgin English) but also had an overall positive attitude for working on languages.

Collection of speaker metadata
In order to make sure there was sufficient sociolinguistic context for the collected data, all the participants had to ideally be interviewed for information relating to their multilingual repertoire and

competence, life history (especially regarding their mobility and social networks), and language ideologies. This information formed the basis of the speaker metadata used in the project, and it was, by necessity, more extensive than the kind of speaker metadata typically collected for projects focusing on a single language (which are often limited to macro-sociological factors such as age, gender, occupation, and degree of schooling, see Di Carlo forthc. for more details).

Direct observation (example of daily life in the field)
In order to better understand the daily life of individuals in Lower Fungom, residents were visited, observed, and interacted with in an informal way, and field notes were produced. This kind of observation was unstructured, and typically only recorded via field notes. This activity, guided by an observation protocol, produced a handwritten, semi-structured field journal.

Recording speech events
Based on the principles determined through corpus theorization (see Section 6.2.1), audio and video recordings were made with a focus on natural conversations among Lower Fungom multilinguals and on events in which relationships could be observed between multilingualism and supernatural beliefs, the desire to mimic others, the desire to exclude select people from communication, the need to obtain a favor, and the need to assert one's gender. Methods used to record these day-to-day interactions without undue influence and by respecting people's privacy included (i) asking the consultant to visibly wear an audio recorder through several hours during a market day (see also Connell 2009:138 on this method) and (ii) extensive field notes about the situational contexts where these interactions took place.

In order to record these day-to-day natural conversations without undue influence of the researcher and also because such conversations may occur in places where it is difficult for the researcher to be present, an audio recorder was worn by a consultant. This not only provided ample data but also minimized the observer's paradox without raising ethical concerns, as the recorder hung from the consultant's neck in a very visible way and recordings either took place in public spaces or the recorded participants were made aware of the presence of the recorder (see also Connell 2009:138 on this method). In all other cases, video recordings were made.

Analyzing speech events
Speech events were analyzed by listening to the recordings with multiple consultants repeating what they and other participants said. The transcription method used did not aim at phonological accuracy but, rather, at overall accessible representations of the sounds that the researcher heard. This was seen to be adequate for the kind of sociolinguistic analyses that were being undertaken.

After transcription had taken place, consultants also translated the content of the recordings into Cameroon Pidgin English (CPE), a language that the researcher spoke fluently. Where possible, some degree of word-by-word translation was undertaken, but this was not prioritized in a way which would prevent analyzing an adequate amount of data for less fine-grained linguistic behaviors.

After the primary transcription and translation was made with one consultant (typically one of the participants recorded) another consultant (typically not among the recorded participants) was asked to listen to the recording, repeat all the sentences, and provide a free translation in CPE. We refer to these consultants as "judges", since they help verify the information provided by the first consultant and the accuracy of their translations. This was seen as a way to counterbalance concerns raised by the fact that the researcher is not an expert in all of the languages present in the recordings. The final annotations reflected a critical "merge" of work sessions with all participants and judges, and some of the most important resulting information for the research was what language each speaker was using at any given time.

During data collection, metadata on the situational context of the speech event was also collected, e.g., how the participants were socially connected to each other, aspects of the locations of the interaction (especially important when video was not available), etc.

A.3 An example of comprehensive speaker's metadata: "Ja'elle"

A.3.1 Introductory remarks

Data for Ojong Diba's PhD thesis was collected from three residents of Lower Fungom who were referred to as major consultants, as well as from seventeen persons with whom they interacted. All of these individuals were multilingual, though to different degrees. The major consultants had repertoires of no less than six out of the thirteen lects (i.e., named languages) spoken in Lower Fungom, in addition to other named languages such as English, Cameroonian Pidgin English, and French.

Sociolinguistic interview guides (see section B3 below) were used for data collection of Ojong Diba's study. The semi-structured interview guides used were made up of loosely written questions which directed the conversation towards the topics and issues of interest to the research. The guides which were used with both major and minor consultants were composed of three interrelated parts. The first part was made up of questions designed to elicit, in great depth, participants' biographic data that could relate to their reported rates of multilingual competence. We included questions about (1) the various names the participant had (see Di Carlo and Good 2014 for a discussion of the significance of the Lower Fungom naming system for people's multilingualism), and (2) the provenance of participants' relatives, from their parents and spouses up to grandparents. This information was relevant to the understanding of some key portions of what we termed "indexical space" in this article (see Section 3.3 of the article).

The second part of the guide sought to produce a list of all the languages/varieties (lects) in which the consultants claimed competence. It also comprised a part which allowed the consultants to grade themselves on their competence in each of the lects they had reported. They were also asked about their acquisition patterns.

The third and final part aimed at gaining insights into the ideas or practices that the respondents associated with each of the lects they professed to be able to understand or speak. We asked the respondents questions such as "When do you use language A?" "Why did you not use language B"? "Can you perform X activity using language C?" "Which language do you use when you meet person X and why?". This part permitted the researcher to explore the rationalizations that respondents had with regard to their language practices.

In addition, the interview guide allowed to further probe into the linguistic choices of consultants which we had already observed, and which needed clarification, especially as far as the minor consultants were concerned. Deeper questions varied from participant to participant and depended on their prior linguistic behaviors with one of the major consultants. Sometimes, based on Ojong Diba's immersive observations and previous recordings, the participants were presented with previously observed scenarios during which the participant exhibited a particular behavior we wanted to probe. This usually refreshed the memories of the participants or encouraged them to go into more details. The interview guide thus enabled the researcher to engage in a free-flowing conversational exchange with the consultants. These one-on-one conversational exchanges lasted no less than thirty minutes and, overall, they allowed to gain a better understanding of the communicative behaviors and beliefs of both the major and minor consultants.

In Ojong Diba's thesis (2019), and in this article, pseudonyms were used instead of the original names of the consultants. The consultants did not have a problem with their real names being used; the decision was entirely that of the researcher in order to prevent any possible unforeseen issues connected with the publication of any of the contents of the speech data found in the corpus should these be relatable to any of the study collaborators.

Below is an example of data obtained from a major consultant through the use of the semi-structured interview guide. Worthy of note, this process of obtaining data was done for all the consultants used for the study; three major and seventeen minor consultants.

A.3.2 Ja'elle's metadata

A.3.2.1 General sociolinguistic profile

Ja'elle was born in Buu in 1981 to a father from Buu and a mother from Mufu. She married a man from Isu who, at the time of the fieldwork, was the Chief of the Health Centre in Abar. She reported being able to speak or understand more than thirteen languages, all the languages and dialects of Lower Fungom plus English [eng, stan1293], French [fra, stan1290], CPE [wes, came1254], and Isu [isu, isum1240] to varying degrees.[1] She reported native competence in (i) Buu, Mundabli, and Mufu [boe, mund1238], all of which are currently classified as a single language in reference sources, though Buu is better viewed as a distinct language from the other two varieties, (ii) Abar, Missong, and Munken, three varieties of Mungbam [mij, abar1238], (iii) Fang [fak, fang1248], (iv) Koshin [kid, kosh1246], and (v) CPE. She reported being fluent in (i) the variety of Naki [mff, naki1238] spoken in the Lower Fungom village of Mashi, (ii) Kung [kfl, kung1260], and (iii) English. She reported limited command of Ajumbu [muc, mbuu1238] and French, and that she could understand all of the remaining lects of Lower Fungom. Her husband spoke Isu, in addition to English and CPE, and she learned Isu following local customs (see below). They had five children with whom Ja'elle spoke only Buu. Her mother, Mama M.F., reportedly spoke Mufu, Mungaka [mhk, mung1266], Abar, Missong, Buu and CPE. However, Mama M.F. only spoke Buu, her husband's language, with Ja'elle. Ja'elle's father reportedly spoke Buu, Fang, Koshin, Missong and CPE. Notice that her father did not speak Mufu, his wife's mother tongue.

A sociolinguistic interview revealed that Ja'elle went to primary school, attended classes one and two at Abar center. For classes three to seven, she was sent to Ekok (Eyumodjock subdivision, Manyu division, South West Region, Cameroon) and lived with her uncle (her father's youngest brother, who was a policeman), who was from Buu. She came back home during summer holidays and then completed Forms one to three in Wum, as there was no secondary school in Abar at the time. She spent some of her holidays back home. During one of these holidays, in 1993, as she was to go to Form four, she became pregnant. During this period, she lived with a former Cameroonian Member of Parliament, honorable Nkangkolo who was also a first cousin of her father. Six months after she gave birth, she travelled back to Ekok where she began to study tailoring. After a while, the uncle who was serving as her guardian had marital difficulties in his polygamous home and also lacked money to pay for her studies. Because of this, Ja'elle had no choice but to do work as a trader. She travelled to Onitsha in Nigeria and bought kitchenware, such as plates, which she sold in Abar. Soon afterwards, she returned permanently to Abar. At the time of our research, she had five children from her previous relationship, having just lost the sixth child that she had with her new husband.

A.3.2.2 Contexts of acquisition

Table A.1 provides an overview of Ja'elle's linguistic repertoire, indicating the languages she knows, her reported degree of competence, and a brief description of how she acquired it. The reported competence number uses the following scale: 0–can neither understand nor speak the variety; 1–can understand the language a little, 2–can understand the language but cannot speak it, 3–can understand the language and speak it a little, 4–can understand the language and speak it well, 5–is fluent in the language. This five-point scale reflects locally salient patterns of characterizing language competence in the Lower Fungom area and other nearby parts of Cameroon. The language names in the table refer to named linguistic varieties in Lower Fungom rather than scholarly linguistic classifications. We discuss the information summarized in the table in detail below.

[1] ISO 639-3 codes and Glottocodes (see https://glottolog.org) are included with language names to facilitate identification of the relevant varieties.

Language	Competence	Contexts of acquisition
Abar	5	Attended the only primary school in Lower Fungom at the time, which was located in Abar
Buu	5	Born in Buu, and Buu is her father's primary language
CPE	5	From classmates and when moving around Lower Fungom
Fang	5	Father has close relatives in Fang, and they have a very good relationship, which resulted in constant contact
Koshin	5	Learned as a child while living with an uncle who had a wife from Koshin and lived outside of Lower Fungom
Missong	5	Primary school in Abar was also located near Missong
Mufu	5	Mother is from Mufu
Mundabli	5	Mother has grandparents from Mundabli, which meant constant contact, and it is also similar to mother's language
Munken	5	Grandmother's mother was from Munken, which meant that the grandmother regularly used Munken with her grandchildren
English	4	In school since it was the language of instruction
Kung	4	Had a close friend from Kung while in secondary school in Wum
Mashi	4	Learned while moving around Lower Fungom
Ajumbu	3	Learned in Buu from a friend who got married to a man from Buu and could not speak Buu
Isu	3	Learned as a result of being married to a man from Isu
French	3	Learned from the uncle that she lived with who was a policeman and during time spent in Yaounde, Kribi and Bafia
Biya	2	Lived around people from Biya who encouraged her to join meetings of Lower Fungom people when she was a student in Wum and meet with with other people from Biya
Ngun	2	Learned while a student in Wum when attending meetings of Lower Fungom people there

Table A.1: A summary of Ja'elle's linguistic repertoire and contexts of acquisition

Ja'elle learned Koshin from the uncle with whom she lived in Ekok, due to the fact that one of his wives was from Koshin. With respect to Fang, Ja'elle's father had relatives in Fang village, and, as a result, made many visits to the Fang area. This was facilitated by the fact that Fang neighbors Buu village, allowing for day trips there. Ja'elle claimed that people from Koshin and Fang did not understand Abar. As a result, she learned these languages to speak with people from these villages.

While Ja'elle lived in Wum during her time as a secondary school student, she joined a savings group comprised of people from Biya and Ngun, both of which are varieties of Mungbam. During its meetings, she learned a few words of Biya and Ngun. People from Biya and Ngun could understand Abar. Because of this, when she met them, she would speak in Abar; but if they knew her personally, they would speak to her in Ngun or Biya, and she would respond accordingly. She also asserted that some people in Biya were the descendants of people who were original to Buu but had fled during historical periods of intertribal conflict and settled in Biya. These people could, therefore, understand Buu, she sometimes spoke Buu with people from Biya.

Ja'elle learned another Lower Fungom language, Kung [kfl, kung1260], from classmates from Kung in Wum. Because there was no secondary school in Lower Fungom at that time, most students from Lower Fungom went to secondary school in Wum. (The language of Wum itself is Aghem [agq, aghe1239].) Ja'elle stated that she had not lived with anyone from Biya, Kung or Ngun. Rather, she had picked up vocabulary items from these languages during her time in Wum.

Ja'elle's mother came from Mufu, and her mother had grandparents who were from Mundabli. The similarity between the varieties associated with Mundabli and Mufu allowed her to understand Mundabli. Ja'elle learned Mashi in a non-systematic way. She was not related to anyone from Mashi but used this language whenever she met someone from Mashi who did not understand Abar. Ja'elle reportedly learned it from a woman from Ajumbu who married a man from Buu and moved to Buu village but could hardly speak any Buu. She and Ja'elle later became close friends.

Ja'elle's maternal grandmother was from Munken. Ja'elle said that in Buu maternal families were considered quite important. Accordingly, her grandmother ensured that she used Munken with her grandchildren so as to preserve this relationship. She also used Munken with someone from that area who did not understand Abar.

Ja'elle reported having a strong aptitude for learning new languages without much difficulty. She also reported enjoying learning and using as many languages as possible. In her opinion, knowing several languages largely guaranteed her safety.

A.4. Sociolinguistic Interview Guide

Here we provide a copy of a version of the sociolinguistic interview guide used in the research described in the paper that is slightly updated with regard to the one that was used in Ojong Diba's (2019) study. This interview guide is intended for use during fieldwork in Cameroon and includes content that is customized for this purpose. For instance, the numbers of the language competence scale described in section A.3.2.2 above are characterized using language that would be more readily comprehended by interviewees than standard English would be. The interview guide is divided into three sections: a short initial section to capture basic metadata about the interview, an extensive biographic questionnaire, and a section to gather information on each language a speaker reports knowledge of. The third section contains a repeatable part of the questionnaire to gather information on how knowledge of that language was acquired and the contexts in which the interviewee uses it.

BASIC METADATA OF THE RECORDING	
a - **Researcher**	
b - **Date**	
c - **Audio files**	
d - **Place of interview**	

	BIOGRAPHIC QUESTIONNAIRE
	PERSONAL DETAILS
1	Gender
2	In which year were you born?
3	If you were to be born at home and not at the hospital, which village would have been your birth

	place? Give the name of the village and the quarter of birth.
4	What is your current occupation? If you do more than one job, please list all the jobs that you have done over the past 2 years.
5	Where do you currently reside? Village, quarter, compound
6	What are your names?
7	What is/are the name(s) that your father's family gave you?
8	What is/are the name(s) that your mother's family gave you?
9	What is/are your father's name(s)
10	Do you have any other names given by any other relatives?
11	In which quarters / villages did you live when you were between 0 and 10 years old?
12	In which quarters / villages did you live when you were between 10 and 20 years old?
13	In which quarters / villages did you live when you were between 20 and 30 years old?
14	In which quarters / villages did you live when you were between 30 and 40 years old?
15	In which quarters / villages did you live when you were between 40 and 50 years old?
16	In which quarters / villages did you live when you were between 50 and 60 years old?
17	In which quarters / villages did you live after you were 60 years old?
18	What are the schools that you attended?
19	The last time you were in school, what class were you attending and in which school?
20	If your father was to be born at home, not at the hospital, which village would have been his birth place? Give the name of the village and the quarter of birth.
21	Where has your father spent his life (list all villages / quarters in which the father has spent his life with approximate periods)?
22	Where did your father's mother come from (village and quarter)?
23	Please list all other families / quarters in which your father has blood relations.
24	What level of school education has your father reached?
25	What languages can your father hear or speak? Please list

26	If your mother was to be born at home, not at the hospital, which village would have been her birth place? Give the name of the village and the quarter of birth.
27	Where has your mother spent her life (list all villages / quarters in which the mother has spent her life with approximate periods)
28	Where did your mother's mother come from (village and quarter)?
29	Please list all other families / quarters in which your mother has blood relations.
30	What level of school education has your mother reached?
31	What languages can your mother hear or speak? Please list them;
	SPOUSE(S)
32	If your spouse was to be born at home, not at the hospital, which village would have been his/her birth place? Give the name of the village and the quarter of birth. If you have or have had more than one spouse (polygamous man, widow, widower, divorced), please list the provenance of all your spouses, past and present, and assign a number to each one of them (e.g. spouse 1, spouse 2, etc).
33	What is the name and location of your spouse's father's family? For multiple spouses, list their father's provenances preceded by the spouse's number (see question 32)
34	What is the name and location of your spouse's mother's family? For multiple spouses, list their mother's provenances preceded by the spouse's number (see question 32)
35	What languages can your spouse hear or speak? For multiple spouses, list their languages preceded by the spouse's number (see question 32)
36	How many spouses do you have? What level of school education has your spouse reached? For multiple spouses, list their level of school education preceded by the spouse's number (see question 32).
	OTHER NETWORKS
37	Where do your best friends (not relatives) come from (village & quarter)?
38	Please list the names and locations of all the savings groups (Njangi) in which you are member.
39	Please list all the groups in which you are member, besides families and njangis (e.g. dance groups, churches, village societies, etc). For each group, please also say where it usually meets and where the other members come from.
40	When you are sick and want to rely on traditional medicine, which traditional doctor do you go to? Where are these doctors based?
41	Which year did you leave the village? ………………………... Are you an Internally Displaced Person? Yes or No:……………... Where are the various places you lived in after you left the village? List the names of the village(s), town(s), or city(s) you lived in (including quarters).

KNOWN LANGUAGES

Date…………………………… Place of interview…………………………………………....
Consultant's paternal name ………………..………………………………………………….

42. Do you speak, Abar, Ajumbu, Biya, Buu, Fang, Koshin, Kung, Mashi, Missong, Mufu, Mundabli, Munken, Ngun, Pidgin, English, French, any other languages? Fill competences in the table below:

Language name: **Do you speak / hear?….**	**Degree of competence:** 0 = can neither hear nor speak; 1= hears a bit; 2 = hears but no talk; 3 = talks a bit; 4 = talks well; 5 = fluent
Abar	
Ajumbu	
Biya	
Buu	
Fang	
Koshin	
Kung	
Mashi	
Missong	
Mufu	
Mundabli	
Munken	
Ngun	
Pidgin	
English	
French	
Any Others	

LANGUAGE SHEET – ONE SHEET = ONE LANGUAGE / LECT

Language / lect …………………… Consultant's paternal name …………………………

B1	Language name	
B2	How did you learn it and where?	
B3	When do you use it?	
B4	Are there any special occasions in which you use it? (e.g. prayers, songs, invocations, formulas) **Get details.**	
B5	Do you ever have dreams in this language?	
B6	What are the advantages of knowing this language?	
B7	If you did not know this language, what would be the consequences?	
B8	How do you feel when you use this language (e.g. comfortable, uncomfortable)	
B9	What do you want that people should think (say) about you when you use this language?	

REMARKS (e.g. the interviewee seems shy due to the presence of the husband, the interviewee is perhaps tipsy (need to re-interview the person), etc.)