SP28 (2023), pp. 5–38
http://nflrc.hawaii.edu/ldc
https://hdl.handle.net/10125/74719

The Acquisition Sketch Project

Sketch Acquisition Manual (SAM), Part I: The sketch corpus

Rebecca Defina,[1,2] Shanley E. M. Allen,[3] Lucinda Davidson,[1,2] Birgit Hellwig,[4] Barbara F. Kelly,[1] Evan Kidd [2,5,6]

[1] University of Melbourne, [2] ARC Centre of Excellence for the Dynamics of Language,
[3] University of Kaiserslautern-Landau, [4] University of Cologne,
[5] Max Planck Institute for Psycholinguistics, [6] Australian National University

Abstract

This paper presents the first part of a guide for documenting and describing child language, child-directed language and socialization patterns in diverse languages and cultures. The guide is intended for anyone interested in working across child language and language documentation, including, for example, field linguists and language documenters, community language workers, child language researchers or graduate students. We assume some basic familiarity with language documentation principles and methods, and, based on this, provide step-by-step suggestions for collecting, analyzing and presenting child data. This first part of the guide focuses on constructing a sketch corpus that consists of minimally five hours of annotated and archived data and which documents communicative practices of children between the ages of 2 and 4.

Keywords: language acquisition, language socialization, child language, child-directed language, corpus research

1 Introduction

According to the best estimates, there are just over 7,000 languages currently spoken or signed on the planet (Simons & Fennig 2017). A key characteristic of human languages is their diversity (Evans & Levinson 2009), and so any serious theory of language phenomena must be built upon a representative sample of the world's languages. However, languages are dying at an alarming rate (Evans 2010; Hale et al. 1992), underscoring the urgency of widening the evidential base, which historically has been uneven across subfields of Linguistics. For instance, while fields like linguistic typology presuppose a focus on crosslinguistic diversity, primarily lab-based disciplines like psycholinguistics can only lay claim to studying less than 1% of the world's languages (Norcliffe et al. 2015).

Our focus here is on first language acquisition, which, despite a proud history of crosslinguistic work (e.g. MacWhinney & Bates 1989; Slobin 1985–1997; Berman & Slobin 1994), has data covering only 1–2% of the world's languages (Kidd & Garcia 2022; Lieven &

Stoll 2009). We comprise a group of researchers who all have an interest in minority and endangered languages and their acquisition and maintenance. Over several years we formed an informal working group that is guided by the general question: *how can we tractably study the acquisition of minority and endangered languages in a way that responds to both rapid endangerment and the need to widen the evidential base of the field?* This manual is our attempt at an answer, which builds on the crosslinguistic developmental models of Slobin (1985–1997) and Berman & Slobin (1994), and also takes into account recent work by our contemporary colleagues (see references throughout the manual). We take our central inspiration from Dan Slobin and colleagues' (1967) *Field manual for cross-cultural study of the acquisition of communicative competence*, where they propose "to guide investigators in the collection of comparable cross-linguistic and cross-cultural data on the acquisition of communicative competence" (Slobin et al. 1967: ix).

There are different ways of building upon these past efforts, and scholars have approached this task in various ways. A central approach revolves around the in-depth study of selected languages that are representative of the diversity of the world's languages. (See especially Sabine Stoll's ACQDIV project, Jansco et al. 2020.) Other approaches develop protocols for day-long recordings in small Indigenous languages (e.g. Casillas & Cristia 2019), propose a toolkit for studying basic linguistic phenomena in children acquiring endangered languages (Pye 2021), and/or propose a model for comparative research within closely related languages (Pye 2017). All approaches are concerned with increasing the coverage of languages within acquisition research. They simply focus on different aspects of the challenge. In particular, they deal differently with the issue of data density in child language corpora. Numerous discussions have highlighted the value of dense sampling (e.g. Lieven et al. 2003; Rowland & Fletcher 2006; Tomasello & Stahl 2004). However, building dense corpora comes with a range of challenges. Among other factors, it takes considerable time and is – for most researchers – prohibitively expensive. We find it useful to think in terms of a trade-off between expanding the depth of coverage (i.e. facilitating in-depth studies of a small number of languages) and the breadth of coverage (i.e. facilitating sketch studies of a large number of languages). There is no doubt that in-depth studies of more varied languages are needed. Realistically, however, the number of such studies will continue to remain small. We therefore see a need for developing proposals that aim at broadening coverage.

In a similar vein to Pye (2021), we propose that one productive way to widen the evidential base for language acquisition, while also addressing the broader problems associated with language endangerment, is to take a documentation approach to acquisition. Building on the foundational articles by Hale et al. (1992) and Himmelmann (1998), language documentation developed a theoretical and methodological framework aimed at creating a "lasting, multipurpose record of a language" (Himmelmann 2006a: 1). In the model we present here, we propose an extension of language documentation principles and practices into the context of language acquisition, including especially:

- a collaborative approach to fieldwork;
- with a focus on recording observable spontaneous language use in dynamic, everyday contexts;
- complemented by approaches that generate ethnographic and metalinguistic data;
- resulting in well-annotated corpora;
- that are archived;

- with a view to serving multiple purposes, including community-based and academic uses, as well as contemporary and future uses.

The primary focus of language documentation has been on the adult language. And while documentary recordings sometimes capture the language use of multiple people, including the talk of young children still acquiring the language, documentary linguists have rarely systematically collected or examined this child language in detail. In our approach, we build on the basic tenets of language documentation summarized above and extend them to child language. Specifically, our approach revolves around constructing a small-scale annotated corpus of naturalistic data that documents the language used by and with children of different ages (the 'sketch corpus'), combined with a description of the corpus data (the 'acquisition sketch').

For child language researchers, acquisition sketches and associated corpora represent a potential fount of unexplored data on typologically diverse languages. For language documentation researchers, they constitute a valuable record of an under-documented type of observable linguistic behavior. At the same time, they have the potential to inform language maintenance and revitalization efforts, as they contribute to a better understanding of processes of inter-generational language transmission. They can also help inform language description efforts, as they provide unique insights into the structure of the adult language, for example through the way adults rephrase child utterances. Combining the two fields of language documentation and acquisition research thus has great potential for mutual benefit. Indeed, many documentation researchers have expressed to us that they would welcome the opportunity to utilize and extend their data collection and expand corpora to allow analysis of language across the lifespan, yet they do not know how to go about systematically recording and analyzing children's language.

Accordingly, this manual has been written with several users and stakeholders in mind. The most obvious user is the researcher with an interest in child language development and who works in, or is a member of, a community that lacks any substantial child language data. However, we anticipate that the background of users will vary considerably. Thus, we hope that the manual is equally as useful to a Master's student studying the acquisition of their own dialect of, for example, Igbo or Japanese, as it is to a field linguist working with communities in the Amazon or in Melbourne, and to a community language worker documenting the language used by and with children in support of their community's maintenance and revitalization efforts.

Unfortunately, it is impossible to cater equally for the needs of all users and stakeholders. In this manual, we assume some basic familiarity and experience with language work and/or research (in particular, with the recording and processing of naturalistic data). This assumption allows us to focus on those issues that are specific to working with children. We are aware, though, that not every potential user will have the same background. Wherever possible, we therefore point to readings that cover specific topics in more detail and that will help you gain an overview and/or acquire skills that we presuppose in this manual. More generally, there exists a wealth of information on language documentation and/or linguistic fieldwork informed by a language documentation perspective. Good starting points are Bowern (2015), Gippert et al. (2006), McDonnell et al. (2018), Meakins et al. (2018), and Thieberger (2012). Furthermore, CoLang (Institute on Collaborative Language Research) offers excellent support for different approaches to language work, particularly community-based language work (www.linguisticsociety.org/content/colang-institute-collaborative-research). Our manual is designed with a view to documenting both spoken and signed child language. Unfortunately, there is comparatively much less information on the acquisition of sign languages, and we will not always succeed in addressing the specifics of sign language research. Again, we will sometimes point to further literature in the hope that this will help sign language researchers to ease into the topic of sign language acquisition.

A realistic set of questions to ask for anyone thinking of compiling a sketch corpus and writing an acquisition sketch is: (i) What will the impact of data collection and sketch writing be? and (ii) To whom will the benefits of the sketch flow?

With respect to (i), we stress that the impact of the sketch is primarily a description of the data at hand, which will inevitably be limited due to the relatively small data set. As mentioned above, our approach is explicitly concerned with increasing the breadth of coverage. Our perspective here is that *any* data on a language for which no or very little data currently exist is better than none, but that any conclusions made from the data must be calibrated accordingly. At the same time, we anticipate that there will be many ways in which the data from individual sketches and combined data sets will have an impact on work in various fields, including that of linguists, psycholinguists, sociologists, anthropologists, and child language experts, among others. Each individual sketch will start the long process of correcting the current skew towards English and other European languages that exists in the field of child language. Sometimes the data will be sufficient (or thereabouts) to make some interesting generalizations, such as the nature of child-directed language or children's early phonological development, and sometimes it will merely identify worthwhile directions for further research. Every sketch will be a triumph in its own right, for both the author(s) and the language community. However, it will be particularly exciting if enough data sets are compiled and sketches written to begin to make crosslinguistic comparisons.

With respect to (ii), in terms of benefits, in the spirit of language documentation, which views the rights of individuals participants and communities as paramount (e.g. Rice 2011), we see the language's speakers and signers as primary beneficiaries. A common lament among adult users of endangered languages is the lack of resources to inform the development of educational materials to help with language maintenance. An acquisition sketch can be a valuable document in this space. For instance, a language might have relatively free word order, but may have been described as having a fixed word order. This can lead to a conundrum for curriculum developers wanting to produce literacy materials. Do they stick to this fixed word order, or do they show more variation with respect to word order? Even the small amount of data collected for a sketch should give a sufficient answer to the question of how much variation in word order is typical in child and caregiver language, and thus whether it makes sense to reflect that in their materials. On another level, the sketch format's focus on documenting natural interaction between and with children is bound to provide insights into a community's typical learning environments and communicative practices that support learning. This knowledge in turn constitutes an invaluable asset for any language maintenance and revitalization program (for a discussion, see especially the Child Language Research and Revitalization Working Group 2017). In Part II (Section 7), we offer suggestions for incorporating community outputs into the sketch format, in the hope that they will eventually feed into practices and materials that will be of long-term benefit to the communities.

A second beneficiary is the researcher and, by implication, the field. The sketch data is rich enough to form the basis for more than one publication, for example different publications focusing on selected high-frequency phenomena in the corpus. While we encourage you to explore options, we have written this manual with a specific type of publication in mind: an 'acquisition sketch' featuring a preliminary description of a wide variety of phenomena of child language and child-directed language. We have organized a peer-reviewed open access publication pipeline for the sketches with *Language Documentation and Conservation*, which includes the archiving of data in a repository of the researcher's choice (see Section 2.2.3). This could be the culmination of a postgraduate thesis, or even the first output that would be the basis for a grant application for a more in-depth study. Thus, the sketch may be the researcher's first or last stop in this space. Our aim in providing a set of guiding principles and data workflow is that the broader field can benefit from the collective efforts of many.

With these points in mind, we now turn to the manual. We present a language documentation and description guide to use in the collection and analysis of child language data in the field, in two parts. Part I guides the researcher towards the creation of the sketch corpus based on a minimum amount of data that could realistically be collected within a short timespan (five hours of data sampled from children 2;0-4;0), thus creating a 'snapshot' of language development of children at specific ages. Part II is a guide for the analysis and reporting of the results, and for the writing of the acquisition sketch.

The following sections lay out the steps needed for the collection (Section 2) and annotation (Section 3) of the sketch corpus.

2 Corpus construction

This section provides an overview of the size and content of the sketch corpus. Section 2.1 explains the structure of the corpus, and Section 2.2 offers practical considerations on data collection.

A preliminary note: We give specific recommendations on corpus construction, which – in an ideal world – should be followed as closely as possible. However, we are mindful of the fact that it is not always feasible to follow these recommendations. For example, in a context of severe language endangerment, you may not have much choice as to what or whom to record. In other contexts, community sensitivities have to take precedence, even if they impact on the data collection. In some cases, we suggest alternative setups, but it is impossible to anticipate all eventualities. We would therefore like to stress that you should treat these recommendations with some flexibility: if it is not feasible to follow a recommendation, then change it. *Any* child data is better than none! If you have any doubts, we strongly encourage you to discuss any issues and anticipated changes with one of the contact persons (see the introduction to this special publication for a list).

2.1 Structure of the sketch corpus

The corpus consists of video recordings of two children at five ages (2;0,[1] 2;6, 3;0, 3;6, 4;0). For each child and each age, a minimum of 60 minutes of naturally occurring interaction should be recorded (resulting in 10 hours of recorded data). Ideally these are consecutive minutes. That is, the camera is recording continuously for 60 minutes. A minimum of 30 minutes of each of these 60 (or more) minute recordings should be transcribed, translated, and glossed (resulting in five hours of processed data). The following sections provide details on the ages and numbers of children (Section 2.1.1), the amount of data to be recorded and processed (Section 2.1.2) and the participants and contents of the recording (Section 2.1.3). Section 2.1.4 outlines the rationale for this setup and includes pointers to further reading.

[1] Ages of children are given in the following format: YEAR;MONTH or YEAR;MONTH.DAY, e.g., 2;0 means an age of 2 years and 0 months (i.e. 2 years exactly), and 2;0.13 means an age of 2 years, 0 months and 13 days.

2.1.1 Ages and number of children

The sketch corpus focuses on children aged between 2;0 and 4;0. In order to capture language development, recording takes place at five different ages: at 2;0, 2;6, 3;0, 3;6 and 4;0, allowing for divergences of ±2 months at each age. This flexibility should make it easier to find children of suitable ages whose families are happy to participate, and/or to include children whose precise ages are unknown. In order to capture variation, two children are to be recorded at each age. There are different possibilities for distributing children over the ages:

- If feasible, record the same two children at each age (i.e. adopt the longitudinal approach illustrated in Table 1). This may be possible if your project has a life-span of two years or more, and if you are planning multiple trips to the field site anyway, if you live in the community, or if local families are willing and able to do the recordings on their own.
- If the longitudinal scenario is not feasible, try to adopt a cross-lagged approach: record the same children at two (or even three) ages. For example, if you plan two fieldtrips one year apart, record the younger ages at your first trip (2;0, 2;6, 3;0), and then record the same children a year later (3;0, 3;6, 4;0).
- Otherwise, adopt a cross-sectional approach: record 10 different children, as illustrated in Table 2. This may be the best approach if you only have one chance to collect data.

Please note: In the longitudinal and cross-lagged approaches, you should aim to record more children at the younger ages, as it is possible that children drop out over time.

Table 1. Sketch corpus: Longitudinal scenario.

Age (±2 months)	2;0	2;6	3;0	3;6	4;0
Child A	30(60)	30(60)	30(60)	30(60)	30(60)
Child B	30(60)	30(60)	30(60)	30(60)	30(60)
Total	60(120)	60(120)	60(120)	60(120)	60(120)

Note. Minutes of transcribed language (suggested recording length in brackets).

Table 2. Sketch corpus: Cross-sectional scenario.

Age (±2 months)	2;0	2;6	3;0	3;6	4;0
Child A	30(60)				
Child B	30(60)				
Child C		30(60)			
Child D		30(60)			
Child E			30(60)		
Child F			30(60)		
Child G				30(60)	
Child H				30(60)	
Child I					30(60)
Child J					30(60)
Total	60(120)	60(120)	60(120)	60(120)	60(120)

Note. Minutes of transcribed language (suggested recording length in brackets).

2.1.2 Amount of data

The sketch corpus distinguishes between amount of data to be recorded (a minimum of 60 minutes per child and age) and amount of data to be processed (a minimum of 30 minutes per child and age), adding up to a recommended minimum total of 10 hours (recorded) or five hours (processed). However, our recommendation is to record as much as you can. This will allow you to select suitable data for processing. In addition, the data constitutes a valuable resource in its own right that can later be processed for follow-up studies. Keep in mind that what you do not record now of a specific child, you will not be able to record later, as the child will have grown older by then.

Three possible recording options are outlined in Table 3. Ideally, try to record each child over the course of an entire day. Such a setup will give you a good record of the learning environment of the children: their daily routines, the activities they engage in, and the type and amount of their interaction with adults and other children. Another option is to record for multiple hours within the same week, so that you capture a fair amount of the language that children encounter and produce at that point in time, with a chance of picking up low-frequency phenomena. Alternatively, record a minimum of 60 minutes.

Table 3. Amounts of data to be recorded and processed.

	Recording (per child and age)	Processing (per child and age)
Ideal	Day-long recording	30 minutes
Alternative	3-5 (or more) hours within same week	30 minutes
Minimum	60 minutes	30 minutes

For each child and age, process 30 minutes of data. Select a section of 30 minutes with a) considerable language from both the child and their interlocutors and b) reasonably good audio/video quality. As flagged above, this should be a continuous recording of 30 minutes. In other words, do not select, for example, multiple five-minute sections that are not connected to each other. Preferably, do not select the beginning of a recording, as it is likely that the child (and interlocutors) will need to accommodate to the recording situation, and the beginning is likely to be exceptionally unrepresentative. Processing is discussed further in Section 3.

2.1.3 Participants and content

Each recording follows a focus child. If possible, select the children and the recording contexts with the following in mind (see Section 2.2.1 for suggestions for identifying such children and contexts):

- Prioritize outgoing talkative children who are willing to be recorded, even if they are slightly outside the targeted age range. Note that this may limit the generalizability of the data to the subset of children who are talkative and outgoing (as well as to those who are of similar age), but it does maximize the amount of data you are likely to get.
- Try to ensure that the children's (language) development is not delayed.
- Attempt to have a mix of child gender.

- In a multilingual community, do not worry about trying to find monolingual children. Record the children who are willing to participate regardless of the number of languages they grow up with. See Part II (Section 3.3) for notes on multilingualism and possible approaches to handling multilingualism within the sketch format.
- Record contexts where talk and interaction with children occurs naturally.
- Include siblings (or other children of all ages) and parents (or other adults) in the recordings, so that the sketch corpus contains information on how different interlocutors engage with children. Do not include too many participants, as this will make the transcription process very difficult.

The rationale behind these recommendations is as follows: to make it possible for the sketch corpus to be of a manageable size, we need to ensure that it contains a sufficient amount of talk from the child (hence the focus on talkative children) and different interlocutors, both adults and other children (hence the focus on multiple participants). Similarly, children whose development is delayed are not ideal participants for such a small study.

2.1.4 Rationale for the setup and further reading

The period between 2;0 and 4;0 is a time when large parts of a language are being acquired (for comprehensive introductions to child language acquisition see Clark 2016; Hoff 2014; Rowland 2014; for more advanced handbooks see Bavin & Naigles 2015; MacWhinney & O'Grady 2015). At 2;0, we are likely to observe a small vocabulary of first words used in utterances of two or more morphemes. At 4;0, the major nuts and bolts of the language will be in place, with children having command over a large vocabulary, producing multi-word utterances and productively[2] using major parts of the morphology. This does not mean that acquisition is complete at age 4;0 (or that acquisition only starts at age 2;0), but the study of language development beyond this point requires different types and amounts of data than we can reasonably hope to collect within a sketch corpus.

Although the overall trajectory of language development is similar across children, individual children differ considerably, and age can only ever be taken as a rough indicator of developmental stage (Bates et al. 1995; Kidd & Donnelley 2020; Lieven 1997). For example, it is entirely possible for a fast learner at age 2;0 and a slow learner at age 2;6 to not be very far apart developmentally. Using proficiency metrics like MLU (mean length of utterance, see Part II, Section 6.3) can be valuable in understanding any differences between children that may not be revealed by age (Brown 1973). There may also be gender-based differences. Research conducted in English-speaking contexts in the USA, for example, has shown small gender effects in language development in favor of girls, which tends to flatten out in the third year of life (e.g. Fenson et al. 1994).

A longitudinal setup circumvents the problem of comparability, as it traces the development for each child individually, and allows us to look at the child's development at different points in time. The sketch corpus aims to approximate such a longitudinal setup,

[2] Productivity sets in when children start analyzing the internal structure of complex utterances (e.g. analyzing *walked* as *walk -ed*), thus enabling them to combine the forms and structures of their language to create new expressions. Note, though, that complex utterances are not necessarily used productively: children may produce them without having analyzed their internal structures, for example as rote-learned forms or as repetitions of an interlocutor's utterance.

ideally covering both girls and boys. However, there are two crucial differences that we need to keep in mind as they impact what we can reasonably say on the basis of the sketch data.

- First, a full longitudinal study employs a much denser sampling rate than our 30(60) minutes across six month intervals, i.e. it will capture considerably more of the language that children encounter and produce.
- Second, a full longitudinal study records the same child(ren) from age 2;0 through to age 4;0. That is, it traces the development of each child individually, not resorting to the cross-lagged approach that is likely to be the most feasible setup in many field sites.

If you are interested in going beyond the sketch approach, we recommend that you consult the literature on constructing acquisition corpora. Good starting points are, for example, the various contributions in Behrens (2008); see also Demuth (1996, 2021); Eisenbeiß (2005).

2.2 Practical considerations of corpus construction

This section provides practical advice on corpus construction: identifying suitable children and contexts (Section 2.2.1), setting up the recording (Section 2.2.2), archiving the data (Section 2.2.3), and attending to ethical considerations (Section 2.2.4).

A preliminary note: We have attempted to keep the sketch corpus to a manageable size, but 10 hours is still a lot of data to record (and 5 hours, to process). To a large extent, you will be able to fall back on best practice methods from language documentation (see Section 1 for some references). Yet be aware that child data comes with its own challenges. In some respects, it is easier to handle than adult data: it can be repetitive and there is a good chance that adults modify their language, producing shorter and more articulate utterances, which are easier to transcribe. In other respects, though, it is more difficult, as it can sometimes be very hard to interpret a child's utterances. In writing up the practical considerations on corpus construction and data processing, we assume a basic familiarity with language documentation methods, and we focus on issues specific to child language.

2.2.1 Getting started: Identifying children and contexts

We recommend recording talkative children without developmental disorders (suspected or diagnosed), interacting in natural settings with a variety of interlocutors. Section 2.1 has laid out the requirements, and this section now offers suggestions for getting started.

The ideal approach for identifying children and contexts involves engaging with community views, interviewing adults and observing children in interaction. Our suggestions build on collaborative research practices in the field of language documentation and anthropological fieldwork. For example, it is highly unlikely that there will be any diagnostic tests for developmental disorders available for your field language(s) (e.g. for a cognitive, social or language impairment). However, you can inquire into the community's views on language development, and you can observe how children interact with others. Be aware that different communities may have different cultural traditions about the appropriateness of non-community members interacting with children; ensure you are sensitive to local traditions.

The purpose of this initial phase of conducting interviews and engaging in observations is twofold. On the one hand, it is to identify suitable children and recording contexts. On the other hand, the purpose is also to collect metadata information on the participating children and their families (see Section 2.2.3) and to identify salient cultural views on language development that will feed into the acquisition sketch (see Part II, Section 4). In the remainder of this section, we provide a set of questions that will help you in these endeavors.[3] Do not feel that you must ask and answer all of them. They are intended to get you started and to point you in possible directions.

First, we recommend discussing the following three points with as many community members as appropriate:

- Which families have children of the right ages? Does the community consider them suitable families for the project (e.g. in terms of social standing, language background and use, etc.)? Do the families themselves have the time and inclination to participate?
- How do people assess and talk about the linguistic and non-linguistic development of children? Which linguistic structures and/or non-linguistic skills are children expected to master at which ages? Which specific children master the identified structures and skills?
- What do children of different ages typically do during various parts of their day? In which contexts do they interact with others, in which not? With whom do they interact?

Once you have identified candidate families (or candidate families have identified themselves to you), address the following points within each family:

- Which structures and skills has the child mastered? Are there any typical non-target-like forms (i.e. forms that differ from the adult target, e.g. *runned* instead of *ran*)? What are the longest utterances the child has produced?
- What does a typical day look like for the child? What does the child do when? With whom does the child interact?
- What is the child's age? In the absence of birth certificates, it may not always be possible to determine the exact age, but we assume this to be a manageable issue, especially given that the sketch format allows for deviations of ±2 months from the target age. Presumably it will be fairly straightforward to determine the order of birth and the relative age of a child compared to other children, thus giving you an idea of approximate ages. Aside from this, the following kinds of sources often prove helpful (if available): baptism records, health cards (especially if the country has a vaccination program with immunizations administered at different ages), and notable events around the time of birth.
- If possible, accompany the family for a longer period (a day or part of a day), and observe the child in interaction. Who talks to the child and in which contexts? Is the child talkative? Does the child initiate interaction and respond to the interlocutors? Try to observe pointing behavior, as this is a good indicator of developmental disorders; if children are not pointing by age 2;0, then their development is likely to be delayed.

The answers to these questions will enable you to identify families that the community considers appropriate for the study, and that are willing and able to participate. Furthermore, the information on children's daily activities will enable you to recognize natural contexts that trigger talk and interaction. Once you have identified such contexts, investigate the possibility of recording them. This includes both technical considerations (e.g. children roaming through the bush or playing in water may present you with insurmountable

[3] For further inspiration, see the guidelines of the Q-Bex project (Quantifying Bilingual Experience) (De Cat et al. 2021): while they focus on multilingualism, many of them are relevant to both monolingual and multilingual contexts.

technical difficulties; see Section 2.2.2) and ethical considerations (e.g. recordings in some contexts may constitute unacceptable intrusions into family life; see Section 2.2.4).

If no suitable contexts emerge, try to explore alternatives: Are there any special occasions that are not part of the daily routine but which involve children? Is it possible to set up such contexts? Are there community members who regularly interact with children and could be recruited as interlocutors (e.g. nurses, midwives or school teachers)? Is it possible to set up alternative contexts that may not be 'natural', but that may nevertheless be successful in triggering talk? This may mean providing materials for the children to use or play with. It could also mean providing food, which tends to be highly effective at putting children at ease. Be aware, though, that food may be consumed in silence and/or that any speech produced with a full mouth may be very challenging to transcribe.

When deciding on recording contexts, you will need to balance two factors. One possibility is to record all data within the same kind of context (e.g. within the family home). This has the advantage that the data is more comparable, as you can exclude variation due to contextual factors. The other possibility is to record in different contexts (e.g. some recordings within the family home, some in the family garden). Here, the positive is that your sketch corpus gives insights into the diversity of children's learning environments. Given that both approaches have their distinct advantages, we do not make any specific recommendation, and only ask you to keep contextual factors in mind when analyzing your sketch corpus.

The above suggestions are intended as pointers to get you started on finding children and contexts. Feel free to flesh them out and alter them as appropriate. You are in the best position to judge their feasibility in the community, and you may even be able to already give (partial) answers to the above questions. Keep in mind that, despite all efforts, you may not have much choice in the end. As always in fieldwork settings, it may be the participants who approach and select you, not you who selects the participants.

2.2.2 Recording setup

The following suggestions regarding the setup of recording sessions are to optimize both the naturalness of the data captured as well as its audio/visual quality. Given that the aim is to collect naturalistic data, the recording environment requires high ecological validity (Eisenbeiß 2010); it ought to be a real-life situation or at least resemble real-life situations as much as possible. While this is relevant at any age, it is especially important when recording young children, who, when ill-at-ease, tend to produce very little language. One way to try to minimize the unfamiliarity of the recording context is to record people in natural groupings. Other ways to foster ecological validity are to record in locations that the families suggest, and to facilitate children's engagement in familiar activities.

Your initial interviews and observations should give you a good idea of suitable situations to record (see Section 2.2.1). This section addresses issues that arise in the recording setup. We recommend a setup that consists of microphones placed on the children plus a (more or less) stationary video camera. Such a setup allows you to capture natural interaction in good audio/visual quality, but be aware that it necessitates either the use of wireless transmitters or the use of separate audio devices that require post-recording synchronization.

Recording audio data

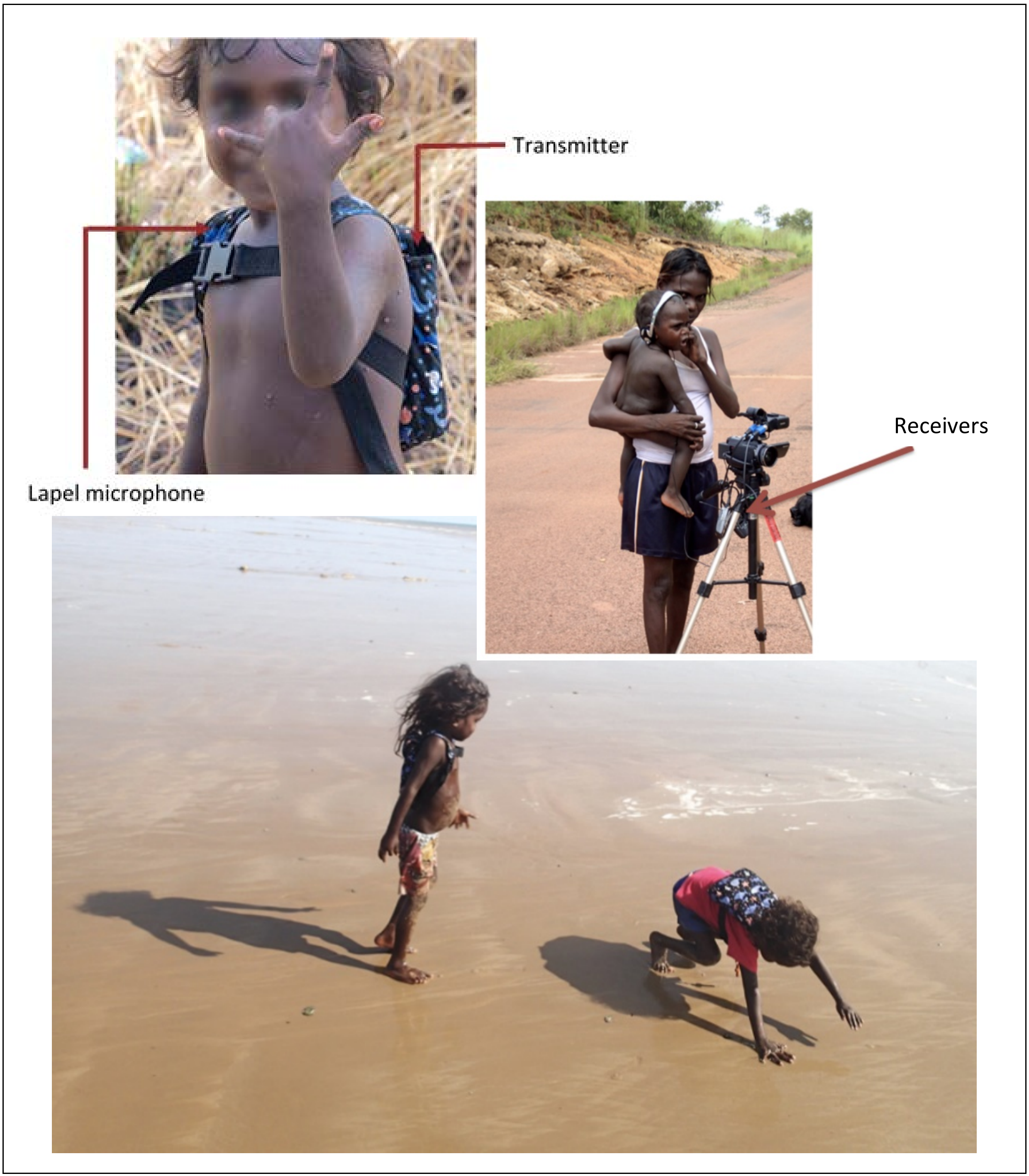
In terms of recording speech production, the best approach depends on the particular context in which you are recording. For example, if recording in a small inside space, a shotgun microphone connected to a video camera is a good option. If recording children in a location where they have the chance to move out of range of a stationary microphone – outdoors or in a large indoor space – we recommend a different method. We suggest keeping a microphone close to each focus child, on their person, in such a way that they are free to run around. A proven approach (illustrated in Figure 1) is to use small bags/backpacks that children wear and which hold a (lightweight) recording device. The ideal position for the microphone is somewhere on the child's front, so as to record speech as clearly as possible. This is crucial for children just starting to produce speech. Another positive of recording children on individual audio devices/channels is that it greatly assists when trying to differentiate individual child voices in preparation for transcription. This method of recording children also picks up speech from other participants when they are nearby. However, if possible, it is a good idea to also record surrounding speech with a shotgun microphone attached to the camera. This will also act as a back-up audio recording for the children should the other equipment fail or be discarded.

As for whether you use a bag positioned on the children's back or one on their front, this is up to the researcher and may involve a degree of trial and error. Researchers have had success with both orientations. Benefits of positioning the bag on the front are that it can be easier to get the microphone closer to the child's mouth and it may be possible to add a small camera to capture still images from the child's perspective, such as the Narrative Clip used in the setup described in Casillas et al. (2020). The downside of positioning the bag in front is that anything on a child's front is likely to cause more distraction and is at greater risk of being meddled with and/or switched off. Whichever approach you take, the most important factors are that the bag/backpack holds the recording device securely, the microphone is positioned fairly close to the child's mouth, the child can move freely while wearing it, and the child cannot remove it easily without help. After all, a masterful design is no help if a child discards the equipment two minutes into the recording. (A plastic buckle clip is a good option if secured across the child's chest.) Equally, if a child finds the equipment cumbersome, the only language they are likely to produce is protestations about the equipment itself.

If recording the children's speech using a wireless transmitter and receiver, place the transmitter in an appropriately-sized bag/backpack and secure the attached lapel microphone inside the bag strap, ideally in such a way that the child cannot easily touch it, e.g. inside a hidden pocket. (The receiver is attached to the video camera.) If using an audio recorder such as a Zoom H1, place it – ideally in a windsock – into the bag/backpack, again in such a way that the child cannot easily touch/remove the device. If using a lapel mic with the Zoom recorder, secure and hide it as mentioned above. If using an audio recorder that is separate from the camera, such as this, try to start the camera and the audio devices recording at the same time or as close as possible (and before the children have the bags put on them). Perform director-style claps within the video frame to aid in syncing the different audio and video streams; these low-tech hacks will help with the alignment of files later. See Meakins et al. (2018) for more detailed advice on equipment and recording.

As always, we also suggest recording in a setting that is as quiet as possible given the field situation. Background noise that humans tend to tune out – such as a dishwasher or TV in the next room, traffic on the street outside, or a running brook or light wind in an outdoor setting – often sounds much louder on a recording and thus interferes with transcription.

Figure 1. Example of recording setup (photos: Lucinda Davidson), showing the location of the transmitter and hidden lapel microphone in the child's backpack, and the receiver plugged into the camera. Children are free to run about while wearing the backpacks.



Recording video data

If your equipment budget allows and if the location is big enough to warrant it, using two video cameras is a great way of capturing the visual component of recordings. Position one camera close enough to participants to capture their nonverbal communication. Keep the other camera further out, capturing the broader scene and allowing people to wander about and remain in view. If recording in a smaller area two cameras might feel oppressive for participants, cause too much of a distraction for children, or simply not be practical. In such cases, use just one camera, and position it close enough to participants to capture nonverbal communication.

As you are dealing with children, and hopefully children who feel fairly comfortable in the recording environment, expect there to be movement. This will likely mean that the camera will have to be moved to keep the children in view. A good way to manage this and simultaneously foster a sense of fun and ease amongst participants with regards to the camera is to encourage an adult or older child to engage in this task. Seeing a familiar person behind the camera tends to help make even very shy children comfortable.

Notes on long recordings

As discussed in Section 2.1.2, we recommend day-long recordings wherever possible. If you do this, or if you record for a period of numerous hours, it is critical to enlist others to operate the camera, and provide basic training to do this. With respect to these longer recordings, it is also recommended to train caregivers to make brief notes about their child's activities. Basic metadata such as 'child asleep', 'child A and B playing with ball' and 'X just arrived' will help enormously when working with the data, as well as helping you initially decide which sections of the recording to look at. See also Casillas & Cristia (2019) for a possible setup using the LENA device, which is a wearable, child-safe audio recorder that can run for up to 16 hours at a time.

2.2.3 Archiving and metadata

The sketch corpus contains valuable data that will be useful beyond the acquisition sketch itself and that should be archived. Several options are available – for example, documentation archives such as ELAR (Endangered Languages Archive) or TLA (The Language Archive), or child language databases such as CHILDES (Child Language Data Exchange System). You are probably already familiar with one or more of the options, or your sketch corpus forms part of a documentation project associated with one of the archives. If this is the case, you should make your own choice and arrangements. Alternatively, we have arranged with the Language Archive Cologne (LAC) to archive sketch corpora.[4]

In any case, you should negotiate with your chosen archive early on in your project. Archives usually have standard setups and procedures for implementing access rights and restrictions and for providing metadata information. While these procedures were developed for the adult language, they are also applicable to child language. Be aware, though, that there are certain characteristics of child data to which you need to pay attention.

First, child data usually cannot be made publicly available in the same way as adult data can. From a legal point of view, many countries do not allow for this possibility at all. From an ethical point of view, the sketch corpus captures unguarded informal day-to-day interaction within a family, giving us glimpses into the personal life and life choices of individuals. It will almost inevitably feature highly personal moments, as well as little dramas, tantrums and tears, and adults' reactions to them. Even if parents give their informed consent to making such data publicly available, they will have to make this decision on behalf of their children. As the children grow older and come of age, they may not agree with the decisions made by their parents.

[4] For more information on the above-mentioned archives, please go to https://www.elararchive.org/ (ELAR), https://archive.mpi.nl/tla/ (TLA), https://childes.talkbank.org/ (CHILDES) and https://lac.uni-koeln.de/ (LAC).

For these reasons, you must carefully think about access rights, access restrictions and the options available for anonymization or pseudonymization. This issue should be discussed early on with your archive, as some archives discourage, or even disallow, implementing access restrictions. To be of use beyond the acquisition sketch itself, discuss at least the following issues with the archive and with the families and communities:

- The possibilities of granting access for specific scientific and/or community-related purposes.
- The possibilities of giving registered users access to parts of the data, for example only to the transcript or to both the transcript and the audio track; possibly even to the video if there are no legal requirements against this.
- The possibilities of anonymizing or pseudonymizing the data, for example replacing names in transcript files with a unique identifier (ID), beeping names in audio files, or blurring the images of children in video files.
- The possibilities of re-negotiating informed consent with the children once they come of age.

When negotiating informed consent, keep in mind that the recordings will probably not only feature the family and their children, but also children of neighbors as well as passers-by. As such you will likely need to discuss these issues with a larger group of people.

Second, we recommend compiling more extensive metadata than is typical for adult corpora, as this information directly impacts the interpretability of the data. Much of this information will be collected in the preparatory stages of corpus construction (see Section 2.2.1), and supplemented with information gained during the transcription process (see Section 3.2.3). Please keep in mind that this is likely to include sensitive information on the child's development, and you need to review carefully which kinds of metadata information can and cannot be made publicly available.

Aim to compile a dossier on each focus child containing the information in Box 1. This information is mostly unstructured, and it is in addition to the structured metadata categories recommended by the archive.

Furthermore, for each recording session, aim to provide the information summarized in Box 2.

Box 1. Metadata: Focus child.

Focus child

(i) Assign an ID or pseudonym, and make sure to use this in publications to protect privacy. The participants may find it fun to suggest their own pseudonyms.

(ii) Name, gender, age (as precisely as possible).

(iii) Any information that you have collected on their linguistic and non-linguistic development, for example their talkativity, their first words, their longest utterances, at which ages they mastered which skills, etc.

(iv) Any information that you have collected on their typical daily routines.

Box 1 (continued). Metadata: Focus child.

Their main interlocutors (even if they do not participate in any of the recordings). This list should minimally include the immediate family (parents, siblings), but it is likely to contain others as well (e.g. grandparents, more distant relatives, neighbors). The goal is to identify and characterize the main interlocutors of the focus children.

(v) ID/pseudonym, name, gender, age.

(vi) Type of relationship to focus child.

(vii) Typical contexts of interaction with focus child.

(viii) Language(s) known, and language(s) typically used with focus child.

Box 2. Metadata: Recording.

(i) The structured metadata recommended by the archive for each participant in a session (e.g. ID/pseudonym, name, gender, age, role in the recording) and the session (e.g. date, location, topic).

(ii) Record the ages of all participating children as precisely as possible and calculate them for each session (if possible in the format YY;MM.DD).

(iii) For each participant, record the type of relationship with the focus child.

(iv) A descriptive account of the context of the recording: setting/location (e.g. “in the kitchen hut, next to the fire”), participants and their contributions (e.g. “the adults talk amongst themselves and only rarely interact with the children, while the children play with each other”), main activities (e.g. “the children play with sticks”), and main topics (e.g. “the children talk about building a house”).

2.2.4 Ethical considerations

Working with children raises a set of ethical concerns that go beyond those that arise when working with adults. From a legal and administrative perspective, it is likely that this kind of research will have to pass through a more complex approval procedure and adhere to stricter requirements. Given that different countries, universities, and communities follow different policies, it is impossible to present an overview here. You will have to inquire about your university’s and community’s procedures. If you have previously worked with adults in the field, you will not have to start from scratch, but can build on your prior experience with ethics boards and the ethical challenges involved in adult fieldwork (see Dwyer 2006; Innes & Debenport 2010; Rice 2006, 2012) and/or can draw on a long-standing relationship with a community that allows you to negotiate and navigate the ethical challenges in appropriate ways. We therefore restrict ourselves to highlighting those challenges that are specific to creating the sketch corpus. For further discussion of the ethics of longform recordings of everyday life in various settings, including example consent forms and participant FAQs, see Cychosz et al. (2020).

One central ethical issue revolves around the selection of the participating children and their families. As discussed in Section 2.1, the format allows for some flexibility, but there remain a number of criteria that could lead to the exclusion of families who would like to participate. Or, conversely, families who meet the requirements may feel community or other pressure to participate even though they would prefer not to. In addition, participating families might draw unwelcome attention from other sections of the community, possibly triggering feelings of envy or suspicion. To some extent, such issues arise in all language documentation work, and we advise addressing them in the following way: widely informing about the project, discussing and making transparent the criteria for in-/exclusion (as well as the purpose of these criteria), creating opportunities for family members to voice their concerns and enabling them to refuse to participate, and finding alternative ways for willing families who do not meet the criteria to be involved (e.g. contributing to interviews on socialization practices, or helping with the transcription of the sketch data).

Another central issue is informed consent, including both the consent to participate in the research and the consent to archive the data. Again, this issue arises in all research contexts. There is a long-standing debate in the language documentation literature about possible ways of keeping both the community and the specific participants informed about the goals and methods of the project, and of how to ensure that any consent to participate and archive is, indeed, *informed* consent. In the context of the sketch corpus, the same issues arise, but are exacerbated for two reasons.

First, the children are minors and any informed consent has to be given on their behalf by their parents or carers. (Children's *assent* to participate in recordings, however, can and should also be sought.) Furthermore, the recordings not only involve the focus children, but very likely a larger number of community members, including the children of others. In other words, you need to consider questions such as the following. Who can give consent for whom? For example, local leaders for the community, parents for their children, or does somebody else need to be involved? Who gives consent for non-focus children such as cousins, or foster children who happen to live with the family, or neighboring children? What happens when the children come of age?

Second, the data potentially intrudes into the personal life of families. This includes the recorded data, as the goal is to capture everyday family life. It also includes the background and metadata information, as it captures sensitive data on the linguistic and non-linguistic development of the focus children and on their social networks. In other words, you will need to ensure that participants are aware of the kinds of data involved, and that they understand the implications of collecting and possibly making them available to others. All questions of access rights and the desirability (or possibility) of anonymization/pseudononymization must be discussed against this background.

The above issues not only arise from the perspective of archiving (see Section 2.2.3), but also from the perspective of the transcription and analysis process, as families may not be happy for other community members to transcribe the data and thus gain insights into their personal lives (see Section 3.1.1).

Finally, ethical challenges arise with respect to the outcomes and benefits for the community. While individual participants should be compensated for their time and efforts (and the precise compensation for particular tasks should be openly communicated), the broader community should also benefit from the work. Again, such considerations are integral parts of any language documentation work. Be aware, though, that potential long-term benefits (such as the role of this research for language revitalization and maintenance efforts, or for developing diagnostic tests for language development or language disorders) are too long-term to be of any immediate benefit and you should also incorporate more short-term outcomes. These outcomes constitute an important part of the overall sketch format, and we offer some suggestions in Part II, Section 7.

3 Data processing

This section addresses issues of data processing: preparation of files for transcription (Section 3.1), transcription and translation (Section 3.2), and further annotation (Section 3.3). Note we use 'annotation' here in the sense from language documentation research covering all additional interpretative information added to the corpus, such as part of speech tagging, addressee information, and general notes. Much of this would be described in child language studies as 'coding', for example one could code a corpus for word order. We have decided to follow documentation terminology for the purposes of the manual.

A few preliminary notes: This section is *not* intended as a step-by-step guide. We assume that many documentation projects will have developed their own preferred workflow and will have made different decisions on issues such as the organization of the transcription process, the choice of software, the structure of the database, or the migration of data through various steps. Our intention is not for you to change your established setup, but to enable you to integrate a child language component into it. We do this by highlighting issues that are specific to processing child language and child-directed language and by exemplifying possible approaches and solutions. In the end, you will have to decide how to best integrate these issues into your setup. As always, please feel free to approach the contact persons to discuss any issues that arise. If you have not yet developed a preferred setup, we strongly recommend that you discuss possibilities with an experienced researcher from your field or region and/or with one of the contact persons.

Most likely, your setup will be based on one of the following:

- A workflow that integrates ELAN (2020) with FieldWorks (2020) or Toolbox (2019).
- A workflow that centers on the CHAT transcription format of CHILDES (MacWhinney 2000; 2021) and the analysis programs in CLAN (2021).

These setups have emerged as best practice within language documentation and child language research, respectively. As such, they serve different audiences, but both are equally standard in their respective fields and both will result in well-annotated corpora. All things being equal, we recommend that you adhere to your established setup.[5]

A second preliminary issue concerns the level of detail. On the one hand, we encourage you to annotate as much as possible, as data collection and processing go hand in hand. The annotation process gives us insights that in turn inform the collection and interpretation of new data; the annotations make the corpora accessible for contemporary academic and community purposes; and they provide the baseline for future annotations (i.e. adding annotations to minimally-annotated data and annotating non-annotated parts of the corpus from scratch). On the other hand, annotation is a time-consuming process, and there will inevitably be a trade-off between the level of detail of the annotations and the overall amount of data that can be collected and processed.

We address this trade-off by adopting principles of language documentation, in particular, a differentiated and incremental approach to annotation. This includes limiting annotations to a subset of the collected data: five hours of annotated data out of a larger amount of collected data. And it includes combining minimal annotations for the entire five hours with more detailed annotations for parts of the five hours and/or specific phenomena.

[5] It is, of course, possible to combine elements from the two setups (e.g. to use CHAT transcription conventions within ELAN). And there are possibilities of converting files between the two setups, but – as always in such cases – the conversion process is not necessarily smooth and error-free. If you are interested in combining elements, please familiarize yourself with the respective other setup and feel free to discuss any specific questions with one of the contact persons.

While we ask you to provide minimal annotations throughout, we leave it up to you to decide on your approach to more detailed annotations: whether to include them at all; and if yes, which phenomena to focus on. In making your decision, please consider the following two points:

- The vitality of the language. In the context of endangered languages, the opportunities for recording data in the future are likely to be limited, and it may be more important to focus on data collection and minimal annotation in order to maximize the amount of data that can be processed.
- Phenomena on which you would like to focus. Part II of this manual provides an overview of topics in child language and child-directed language, and we recommend that you consult this part before deciding on detailed annotations. In particular, ask yourself the following question: are there any topics that you would like to focus on and for which you would need a systematic annotation of (all or part of) the data? If yes, consider a detailed annotation. But in many cases, it may not be necessary to provide such annotations. For example, you may decide on conducting a manual search and annotating the extracted data, or a simple concordance search of the transcript may be enough to find relevant instances. It may be that systematic annotations may only be relevant for some ages (e.g only the youngest age group) or some participants (e.g. only caregivers) or some contexts (e.g. only child-directed language), etc. Depending on the structure of the language, the purposes of the project and/or the research interests, different projects are likely to necessitate different decisions here.

Box 3 summarizes the recommended setup (labeled ‘core’): a segmentation into structural units (e.g. intonation units) and their transcription, translation and analysis on a number of different tiers. Note that this summary reflects the final setup. As discussed in the following sections, we do not recommend annotating simultaneously on all tiers, but rather incrementally adding annotations during various stages of analysis. In each case, it is your decision how much detail to include (e.g. to offer a broader or more narrow transcription of the child utterance). Furthermore, not every utterance needs all tiers (e.g. a general notes tier is likely to only be needed for some utterances); or it might even be the case that a free translation tier is not needed (e.g. because you research a language of wider communication). Overall, we recommend that you attempt to cover the core topics. In addition, you can of course go beyond the minimal setup, and add further analysis tiers (labeled ‘extension’). The following sections discuss these issues in more detail.

Box 3. Key areas of focus in data processing.

Core

(i) Segmentation into structural units, e.g. intonation units (Section 3.1.2)

(ii) Transcription, translation and analysis tiers:

- Transcript of actual utterance (Sections 3.1.3, 3.2)
- Adults’ interpretation of non-target-like utterances (Sections 3.1.3, 3.2)
- Free translation (Section 3.2)
- Morphemic analysis (Section 3.3)
- Addressee of utterance (Section 3.1.3)
- General notes (Sections 3.2, 3.3)

Box 3 (continued). Key areas of focus in data processing.

Extension

(iii) Additional tiers:

Analysis of selected phenomena (Section 3.3)

3.1 Preparing files and transcription sessions

When transcribing the language of young children, some small preparations can have a considerable impact on the amount of work achieved in the transcription session. In order to maximize the amount and the quality of work you achieve, engage with the data beforehand: agree on who does the transcription (Section 3.1.1), segment the files (Section 3.1.2), and decide on the tiers (Section 3.1.3).

3.1.1 Transcribers

Regardless of the setup of your documentation project and the composition of your team, you will need transcribers, i.e. people doing the transcription and translation. This could be you as the project leader working independently on your own language, it could be community members transcribing on their own, or it could be community members transcribing together with you as an outside researcher. In any case, keep in mind that children's language is idiosyncratic, particularly early on in their linguistic development. The ideal transcribers are therefore often the primary caregivers, or at least people who spend considerable time with the child. It is these people who will be the experts on the individual child's articulation, their daily routines, normal contexts, their interests, and their current abilities. Insight into these aspects of a child's life will aid the transcription process greatly. Even if you work on your own language, it will be useful (and sometimes necessary) to include the children's families in the transcription process.

Some of us also have had good experience with involving children in the transcription work, especially when transcribing early language:

- Older children can be extremely helpful in this respect. Often they understand what another child produces considerably better than an adult can. If they cannot decipher what the younger child produces, they may be able to provide a clearer version of the utterance.
- Depending on the child's age, it may also be possible to play the recording back to the child and encourage them to repeat the utterance. This does not always result in a clear reproduction but can sometimes provide valuable information for utterances that are hard to understand.

Engaging children in the transcription often works well when they are present during the transcription sessions and can be called on by the transcribers as they wish. Do not be disappointed if this does not work out and the children cannot decipher the utterances either, but we consider it worth a try.

In any case, discuss the selection of potential transcribers with the family. The family will know who is familiar with the child and thus likely to be able to interpret the child's utterances. They may also not want others to gain insights into their family life.

3.1.2 Segmentation

As discussed in Section 2.1.2, you should aim to process continuous chunks of 30 minutes of recording per child and age. The first step therefore is to identify 30-minute sections that contain the most language produced by the child and the interlocutors. Be aware that ascertaining where there is rich data in a recording involves more than looking for dense audio signals. These sections may well contain considerable child language, but equally it could be talk produced by adults only, or a stretch of crying or other non-speech sounds from children. If you have large amounts of recorded data, it is nevertheless useful to scan for dense audio signals, but once you have identified them, it will be unavoidable that you listen to them to identify sections suitable for transcription.

Once you have identified the relevant sections, they need to be segmented. Ultimately, these segments should reflect structural units of language, such as intonation units or utterances. Several of the analyses proposed in Part II include measurements such as "X number of words per unit of language", thus making it necessary to define the "unit of language" on the basis of clear criteria. The segmentation into intonation units has emerged as standard practice within language documentation (see Himmelmann 2006b), while the segmentation into utterances has emerged as standard practice in child language research (see MacWhinney 2021). Feel free to use either of these options or to segment into another structural unit of your choice, as long as the same strategy is followed consistently throughout. When segmenting child language, be aware that the prosodic and/or syntactic cues that delimit the structural units of the adult language are still developing. Especially for the youngest age group(s), it will not always be easy to decide whether two words in a row should be included within the same structural unit or considered separate units. For example, a repetition of words could count as a single segment (e.g. the child making repeated attempts to produce a single word) or separate segments (e.g. the child insisting on their message). The CHAT manual provides an excellent and accessible discussion of such cases (MacWhinney 2021: chapter 9), and we strongly recommend that you consult the manual, regardless of whether or not you plan to use the CHAT transcription conventions. We also recommend recording your criteria for segmentation within the metadata of the corpus.

While the final segmentation will reflect structural units, different projects will arrive at this goal by different means. For example, if you prepare files for language workers to transcribe, it may be useful to segment beforehand and to exclude those segments that can easily be added later (e.g. repetitions; see below). To avoid any misunderstanding: the excluded segments have to be added eventually, but excluding them from the first-pass transcription reduces the burden on the transcribers. Especially when working with a minority language, it is likely that transcribers will be in short supply, and you will have to think carefully about reducing their workload. By contrast, if you transcribe your own language and/or can rely on larger numbers of transcribers, it may be best to work with the final segmentation from the start, and/or to combine segmentation and transcription within a single step.

Regardless of your workflow, please take the following peculiarities of child language and child-directed language into account:

- Child language tends to be repetitive. For example, a child may repeat a single word or phrase over and over again – sometimes clearly and easily identifiable, and sometimes less clearly and hard to understand. It is useful to identify such repetitions beforehand, so that transcribers do not have to attend to each individual repetition, as this is most likely to either cause fatigue (as each new segment will be identical to the previous one) and/or embarrassment (as the less-clear repetitions may not be understandable on their own). Instead, consider either segmenting only the clearest repetition(s) or including the entire sequence within a single segment. This approach increases the chances of understanding the child's utterances. Any excluded segments can then be added by you or others (e.g. research assistants) at a later stage. Child-directed language also tends to be repetitive, albeit clearly articulated. Again, it is useful to identify and exclude such repetitions beforehand to minimize tedium.

- Child language is not always target-like and may be hard to interpret even for adults who are familiar with the child. Often transcribers will need to resort to contextual information to help interpret an utterance, for example pointing/gesture or eye gaze, the physical environment, response patterns, etc.

Given the above peculiarities, there is a clash between two principles. For transcription purposes, the segments should be long, including, for example, several repetitions and/or contextual information within a single segment. For analysis purposes, by contrast, the segments should be short and reflect a structural unit of language. There are different ways of handling this clash. For example, in our Qaqet sketch corpus, a two-tiered segmentation process was adopted: an initial segmentation into larger units for transcription purposes (as illustrated in Figure 2a, which shows a 4-second long segment), and a later re-segmentation into intonation units (shown in Figure 2b, where this segment was broken up). An alternative possibility would be a segmentation into structural units from the beginning (i.e. like in Figure 2b), but then playing several segments and/or larger chunks during the transcription process.

Figure 2a. Initial segmentation in the Qaqet corpus (for first-pass transcription).

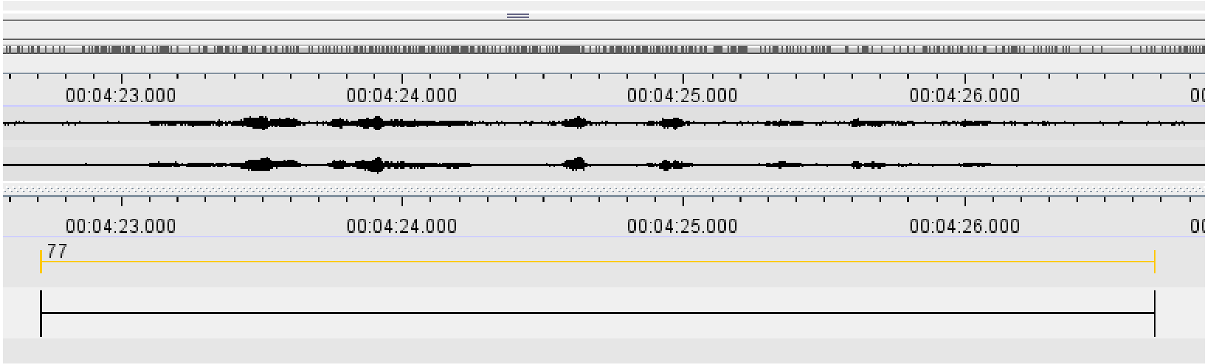
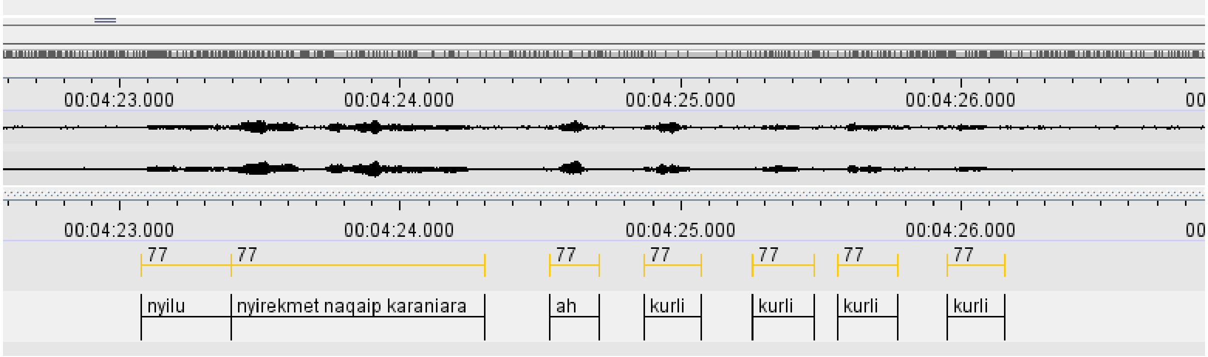


Figure 2b. Final segmentation in the Qaqet corpus (intonation units).

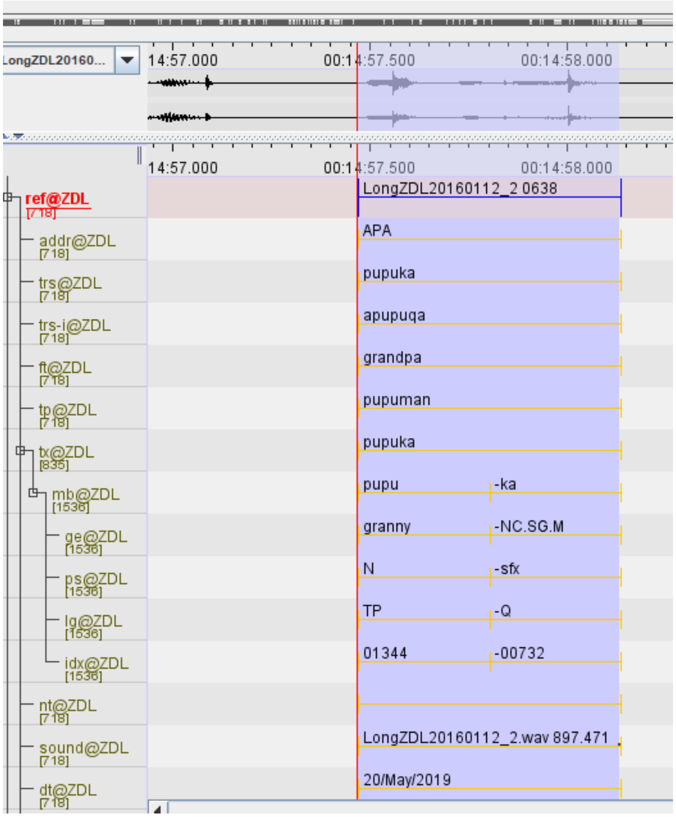


3.1.3 Tiers

As a further preparatory step, you need to decide on the types of tiers needed. Of course, it is always possible to add tiers incrementally as annotation and analysis progress, but experience shows that it is never wrong to plan ahead. The precise number and types of tiers depends on what you intend to focus on in your research, just as it does when working on language data produced by interlocutors of any age. We recommend following standard practice within language documentation, i.e. minimally including a transcript and a free translation tier (see Section 3.2) as well as the tiers needed for a morphemic analysis (morphemic breakdown, gloss and part of speech; see Section 3.3). You may want to include additional tiers for specific analyses (also addressed in Section 3.3), as well as a number of tiers that are specific to dealing with child language and child-directed language (introduced below). Some of the tiers are needed for a first-pass transcription (transcript and translation tiers), others are very useful during that stage (interpretation and/or general notes tiers; possibly also an addressee tier), and yet others will only become relevant during later analysis stages. Again, projects will differ in how they organize their workflow and will thus make different decisions.

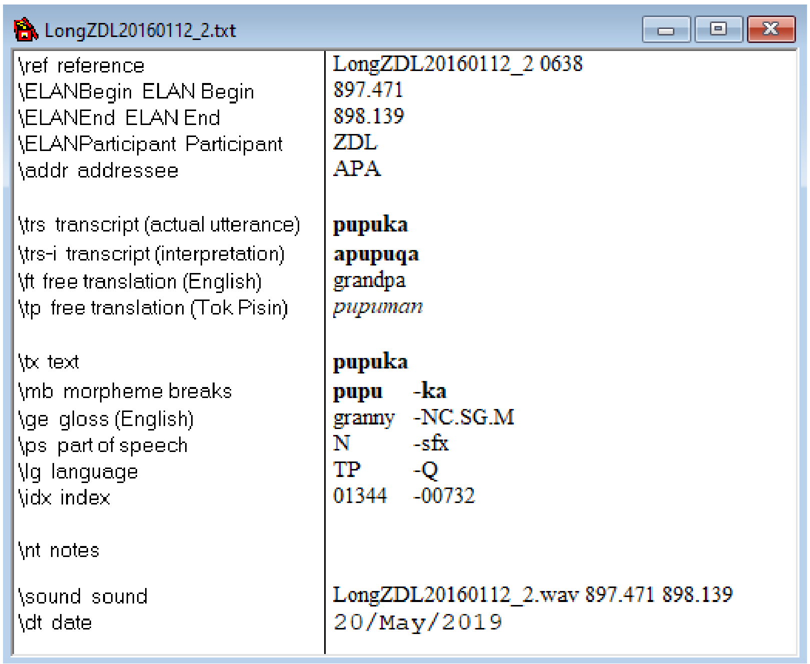
By way of an example, Figure 3 illustrates the minimal tier setup used in the Qaqet corpus (excluding specialized annotation tiers), displaying an annotated child utterance in ELAN (Figure 3a), Toolbox (Figure 3b) and CHAT (Figure 3c).

Figure 3a. Tier setup in the Qaqet corpus: ELAN.



ref@ZDL [718]	LongZDL20160112_2 0638	
addr@ZDL [718]	APA	
trs@ZDL [718]	pupuka	
trs-i@ZDL [718]	apupuqa	
ft@ZDL [718]	grandpa	
tp@ZDL [718]	pupuman	
tx@ZDL [835]	pupuka	
mb@ZDL [1536]	pupu	-ka
ge@ZDL [1536]	granny	-NC.SG.M
ps@ZDL [1536]	N	-sfx
lg@ZDL [1536]	TP	-Q
idx@ZDL [1536]	01344	-00732
nt@ZDL [718]		
sound@ZDL [718]	LongZDL20160112_2.wav 897.471	
dt@ZDL [718]	20/May/2019	

Figure 3b. Tier setup in the Qaqet corpus: Toolbox.



LongZDL20160112_2.txt

\ref reference	LongZDL20160112_2 0638	
\ELANBegin ELAN Begin	897.471	
\ELANEnd ELAN End	898.139	
\ELANParticipant Participant	ZDL	
\addr addressee	APA	
\trs transcript (actual utterance)	**pupuka**	
\trs-i transcript (interpretation)	**apupuqa**	
\ft free translation (English)	grandpa	
\tp free translation (Tok Pisin)	*pupuman*	
\tx text	**pupuka**	
\mb morpheme breaks	**pupu**	**-ka**
\ge gloss (English)	granny	-NC.SG.M
\ps part of speech	N	-sfx
\lg language	TP	-Q
\idx index	01344	-00732
\nt notes		
\sound sound	LongZDL20160112_2.wav 897.471 898.139	
\dt date	20/May/2019	

Figure 3c. Tier setup in the Qaqet corpus: CHAT.

*ZDL:	pupuka
%int:	apupuqa
%eng:	grandpa
%tkp:	pupuman
%mor:	n|pupu=grandparent&TP+nc|ka=sg&m&Q
%pos:	n+sfx
%add:	APA
%not:	
%snd:	LongZDL20160112_2.wav 897.471 898.139

The remainder of this section introduces two tiers that are specific to child language and are beyond the standard setup used in language documentation: an interpretation tier and an addressee tier.

First, two tiers are needed for the transcription of child language. One tier records the actual utterance of the child – the usual transcription tier, labeled '\trs' in Figures 3a and 3b, and '*ZDL' in Figure 3c (the three-letter code for the child's name). The other tier records the equivalent utterance in adult language – the interpretation tier, labeled '\trs-i' in Figures 3a and 2b, and '%int' in Figure 3c. The adult version is essentially an interpretation of the child's non-target-like production, noted on a separate tier. It is also sometimes considered the 'intended' utterance of the child. Be careful with this term, though, as the intention of the child is not necessarily clear. During the transcription process, you should strive to document both the transcription of the child utterance and the adult version of the utterance (see Section 3.2.2 for more information). In this context, we also recommend adding a tier that allows you to note any analytical difficulties that emerge in this process (labeled '\nt' or '%not' in Figure 3).

Second, in order to analyze child-directed language (see Part II, Section 5), you will need to identify and record the addressee of an utterance (the addressee tier, labeled '\addr' or '%add' in Figure 3). You can either record their pseudonomized ID or their role (e.g. 'mother', 'child', etc.), keeping in mind that an utterance may have more than one addressee, or a non-human addressee (e.g. an animal or a toy), or no addressee (e.g. a child talking to themselves).

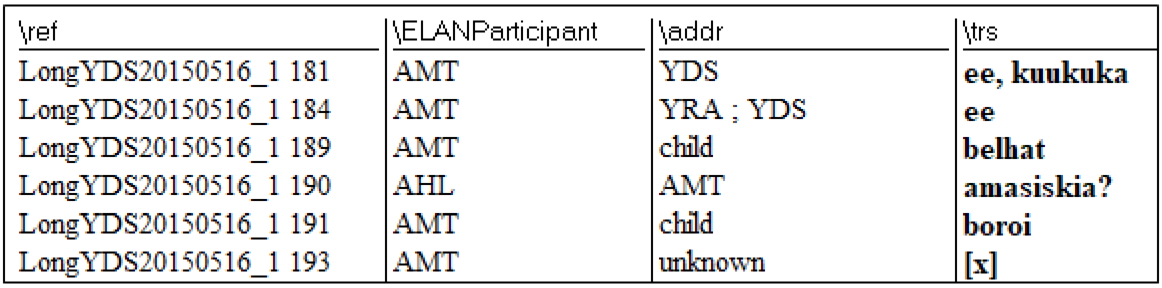
In many cases, content and participant constellation will make it possible to unambiguously identify the addressee(s). In problematic cases, it may be necessary to consult with the transcribers and/or the child's family, and to pay special attention to the following types of cues:

- visual cues, such as body position or gaze direction;
- the use of personal pronouns in transcripts;
- response patterns (i.e. who responds to an utterance).

Be aware, though, that there will always remain a residual group of utterances that are not easily attributable. In some cases, you may be able to determine that it was one of the children who was addressed (without being able to identify a specific child), while other cases will continue to remain unclear. Given that the purpose of annotating for the addressee is to identify child-directed utterances for analysis, we recommend not spending too much time and effort on the unclear cases. In any case, do not be tempted to identify addressees on the basis of the presumed features of child-directed language. For instance, do not identify an utterance as child-directed simply because it is short and high-pitched. Such an approach would lead to circularity: as our investigation aims to identify the features of child-directed language, these features should not be used to identify such utterances in the first place.

By way of an example, Figure 4 illustrates the annotation scheme adopted for the Qaqet sketch corpus. If the addressee is known, their ID is entered (e.g. 'YDS'). In the case of several addressees, all IDs are entered (e.g. 'YRA ; YDS'). If their role (but not their identity) is known, the role is entered (e.g. 'child'). Otherwise, 'unknown' is used. Other codes used but not illustrated are: 'self' (talking to themselves), 'animal' and 'object' (talking to an animal or object).

Figure 4. Annotating for Addressee (example from the Qaqet sketch corpus).



\ref	\ELANParticipant	\addr	\trs
LongYDS20150516_1 181	AMT	YDS	**ee, kuukuka**
LongYDS20150516_1 184	AMT	YRA ; YDS	**ee**
LongYDS20150516_1 189	AMT	child	**belhat**
LongYDS20150516_1 190	AHL	AMT	**amasiskia?**
LongYDS20150516_1 191	AMT	child	**boroi**
LongYDS20150516_1 193	AMT	unknown	**[x]**

3.2 Transcription and translation

While transcription of any language data is a notoriously time-consuming and effortful task, transcribing the language of children, particularly that of young children, presents specific challenges. The aim of the following sections is to help maximize what you achieve in transcription sessions when working with child data. Again, we assume general familiarity with the transcription and translation process, and only highlight the issues specific to child language.

For the translation into a language of wider communication and/or regional languages, please resort to your own preferred setup. We only recommend that you add a free translation in order to make the corpus accessible to wider audiences. Other than that, we have no special recommendations here.

For the transcription, we focus on three topics. Section 3.2.1 addresses the main challenge in transcribing child language: deciphering children's utterances. Section 3.2.2 discusses the two transcription tiers: the transcript of the actual utterance, and the adult interpretation of the child's utterance. Section 3.2.3 highlights the potential of transcription as data collection.

We recommend approaching transcription as an iterative process: that is, returning to the first transcript multiple times, modifying and improving it along the way. For example, it is often useful to exclude repetitions, interjections, hesitations, false starts, and the like from the first-pass transcript. Such exclusions will reduce the burden on the initial transcribers, allowing them to concentrate on the task of deciphering the children's utterances. The other elements can easily be added at a later stage by a second transcriber. The second transcriber can be another community member (in those cases where enough trained transcribers are available), but it can also be an outsider (e.g. an outside researcher, or even a trained research assistant). Furthermore, we have found that our understanding of the children's language improves continuously throughout the annotation and analysis process. In fact, it often becomes necessary to go back to the original audio/video file at later stages, such as when adding another type of annotation or when proposing or verifying an analysis. This process allows us to detect and correct transcription errors – both one-off errors and systematic misinterpretations of children's utterances.

3.2.1 Deciphering utterances

Children's utterances are frequently non-target-like and hard to interpret. We therefore recommend that you engage with transcribers who are familiar with the child and/or who participated in the recording session, and that you transcribe the data as soon as possible after the recording. In addition, the following things may aid the deciphering process.

Consider going through the entire session before starting on the transcription. This will not only give you an impression of the overall context and types of interactions that take place, but will also help you tune into the child's articulation and prosody. We have found that this impressionistic background knowledge greatly facilitates subsequent transcription.

Repetition is a common feature of young children's language, and one that can be of great help when transcribing. The more a child repeats a word or phrase, the more chances there are of being able to decipher what they are producing. For example, the first five iterations of a child's production of a particular word or phrase might be unintelligible, but the sixth might be clear enough to understand. Having understood the sixth iteration, the prior and subsequent productions also become intelligible. Note that these repetitions may not even be contiguous. They often are, but it is just as common that a child is fixated on the same topic over longer stretches of time, and that an utterance much later in a recording may shed light on an earlier utterance. It is therefore often useful to return to undecipherable utterances after the completion of the initial transcript.

Another key to deciphering a child's productions is to observe what is happening locally in the child's environment. Look at the verbal and non-verbal reactions of others present following a child's utterance. Another interlocutor may repeat or rephrase a child's production themselves, providing a more accessible version for the transcriber. Similarly, someone passing an apple to a child might illuminate a prior request for food. Also note any other salient features of the recording context. For example, audible bird song might make clear a child's labelling of 'bird'.

While there are several ways of facilitating the transcription process such as those just described, it will likely be impossible to transcribe all of a child's language productions with certainty. Some utterances will remain unclear to everybody, no matter how familiar they are with that particular child's language or how high-quality the audio/video recording. You will need to develop a tolerance for uncertainty and aim for a balance between quality and practicality. We recommend that you consider adopting the following set of principles:

- decide on the number of times to review an utterance before deciding to either go with a best guess or decide it is unintelligible;
- decide on criteria and conventions for transcribing a best guess if you are not 100% sure (e.g. you could transcribe such segments as [xxx] and add a best guess on another tier, you could transcribe such segments as [yyy] and provide a phonetic transcription on another tier, you could transcribe your best guess on the main tier and flag the uncertainty on another tier);
- decide on criteria for when to ask another person to provide a second opinion.

3.2.2 Transcription and the adult interpretation

When transcribing child language, keep in mind that you need two tiers: one to transcribe the actual utterances, and one to transcribe the adults' interpretation (see Section 3.1.3). In a way, this setup is comparable to the separation of the transcript from the cleaned-up (or 'edited') version of the transcript (where hesitations, errors etc. are taken out). This same approach is taken in many language documentation projects to create community materials (e.g. Mosel 2006).

For the transcription of the actual utterance, pay special attention to the following two issues:

First, the pronunciation of young children may differ considerably from that of adults, and it is unlikely that the practical orthography will suffice to render the language of young children. It is, of course, always possible to aim for a narrow phonetic transcription using IPA, but this is a very time-consuming task, and the audio quality may not be sufficient. For the purposes of the sketch format, we have had good experience with using an adapted version of the existing practical orthography, adding a few more graphemes that allowed us to capture non-adult-like articulations in children's language, while still being easy to implement. However, as always, the decision depends on your interests and goals: if you are interested in a more thorough investigation of phonetics and phonology, your transcript will need to be more precise and probably necessitate the use of IPA.

Second, regardless of your orthographic decisions, your transcript should be as faithful as possible to the actual utterance, including false starts, hesitations, filled pauses, and so on, as they play a key role in the analysis of child-directed language and provide valuable clues to understanding child language. Part II (Sections 5.3 and 6.1) also lists a number of prosodic phenomena of interest (e.g. pitch contours, F0, durations and pauses). However, rigorous transcription of these is very time-consuming and clearly beyond the scope of a first-past transcript. But we do ask you to consider annotating part of the data for these phenomena at later stages.

During the transcription process, you are likely to encounter the following two pitfalls:

First, transcribing young children's language often involves a degree of interpretation, even from their closest caregivers. When caregivers join the project team as transcribers, they may feel particular pressure to understand all utterances that a particular child produces, as they are the experts on this child and their abilities. Until you are familiar with each individual child's way of speaking or signing, it can be difficult to judge how reliable/fanciful an adult's interpretation of children's language is. One way to limit the number of aspirational interpretations is to assure caregivers that not understanding is a valid option. It is okay if they are not sure of what a child is producing. Following this, if a caregiver expresses doubts about an interpretation of what a child has produced, make sure to note this, perhaps on a 'notes' tier.

Second, transcribers are often tempted to provide the 'proper' version of what the child is producing, as opposed to what the child actually produces. In practice, it can be challenging to extract this unmodified information. A promising way of handling this issue is to simultaneously work on the transcript and the interpretation tiers. Allowing the caregiver to provide the 'proper' version, and recording this form on your interpretation tier, may mitigate their reticence to relay a child's 'mistakes' or 'baby talk'. Seeing that you have noted down the 'correct' version, caregivers may be more comfortable providing the child's version. The main pitfall is likely to be that transcribers elaborate too much, adding more material than necessary. We therefore recommend that you plan for enough time to discuss the proposed adult version as well as possible alternatives, possibly even separating between a minimal adult version (i.e. only correcting non-target-like forms) and a more elaborate adult version (e.g. including more complex structures, or stylistic preferences).

3.2.3 Transcription as data collection

Transcription sessions are a great way of collecting data that will enrich your metadata (see Section 2.2.3) and inform your description of socialization practices (see Part II, Section 4).

They provide an excellent opportunity to collect further information about individual children's language, as well as ethnographic information about the child and their family. As such, these sessions provide an appropriate, and in a way natural, context for gaining information that you would want in your metadata for a focus child.

At the same time, transcription sessions also provide a chance to talk beyond the particular children in the recordings and can spark a conversation about language development of children more generally, as well as ideas, beliefs, and practices with respect to language learning that exist in the community and culture. For example, the creation of the adult interpretation of the child's utterance greatly facilitates conversation about common 'mistakes' that children are observed making at certain ages/developmental stages. Observing particular children using the local lingua franca in the recordings could lead to a more general discussion of patterns of multilingualism and/or beliefs about why some children do not use the local language. Please refer to Section 2.2.3 and Part II (Section 4) for topics of interest.

3.3 Beyond transcription

Aside from transcription and translation, you will have to decide on any further types of annotation. Most likely you will first complete a (basic) transcript and translation, and then incrementally add annotations during various analysis stages. As highlighted in the introduction to Section 3, the types of annotations will depend on various factors (e.g. vitality of the language, structure of the language, purpose of the project, research interests), and the sketch approach allows for considerable flexibility here. However, we recommend adding minimally a morphemic analysis ('interlinearization') of the actual utterance (i.e. not of the adult interpretation) to make the corpus i) searchable and ii) accessible to a larger audience, including future audiences.[6] We assume that the morphemic analysis is added to the original transcript (e.g. in a Toolbox-style or CHILDES setup), but please draw on your prior experience and resort to your preferred setup. Again, there are peculiarities of child language that you should take into consideration.

Specifically, given the many non-target-like realizations, it is not always straightforward to decide whether or not a given form should be analyzed as a morpheme and, if so, as which morpheme. Such analytical problems arise when morphemes are only partially realized, as illustrated in the example from the Qaqet sketch corpus below. In this case, we happen to know the target form, because the child makes several attempts at repeating the utterance of his interlocutor. Knowing the target allows us to analyze the child's utterance as containing the morphemes *papa* 'papa', *-iam* 'dual masculine (noun class suffix)' and *a* 'distal (demonstrative)'. However, out of context, the realization [am] would have suggested the noun class suffix *-am* 'dual reduced' (which can be used with humans, but has negative connotations). Should we better analyze [am] as representing *-am*? Furthermore, even knowing the target makes it hard to interpret the initial nasal [m], as it is phonetically very different from the target *gu*. It is much more similar to the articles *ma* and *ama* (which occur in the same slot). Should we better analyze it as representing one of the articles, and if, yes, which one? Or as a filler syllable, i.e. as a syllable that appears in the slot where a particular class of morpheme would normally appear (see Peters 2001; see Part II, Section 6.3)?

(1) child (1;11): *mpapama* target: *gupapaiama*
m=papa-iam=a *gu=papa-iam=a*
DET=papa-DU.M=DIST 1SG.POSS=papa-DU.M=DIST
'two parents' 'my two parents'

There is no easy solution to the questions raised above, but here we formulate some rough guidelines that may help you decide.

- In cases of phonetic and structural similarity (or identity) to an existing morpheme, analyze it as this morpheme unless contextual information suggests otherwise. For (1), this means that [am] is analyzed as *-iam* (whereas out of context, it would have been analyzed as *-am*). Furthermore, initial [m] is not analyzed as *gu-* because of its phonetic dissimilarity.
- You may have to adapt your glossing dictionary in order to deal with uncertainty. For (1), this means that initial [m] is analyzed as a general 'determiner'. Qaqet has an obligatory

[6] It is sometimes recommended to do an interlinearization of the adult interpretation (not of the actual utterance). This approach also serves the goal of making the corpus searchable and accessible, and is hence equally appropriate. Our recommendation of interlinearizing the actual utterance is based on our experience that the adult interpretation tends to be more elaborate (i.e. going beyond correcting non-target-like forms). For example, Qaqet transcribers regularly add subordinating conjunctions in contexts where they are not obligatory (but are judged to be better stylistically). As always, you are in the best position to decide on the best approach for your corpus.

pre-nominal determiner slot (that contains various articles and possessor indexes), and initial [m] is taken to fill this slot. But since we are not able to identify the determiner, a new gloss ("DET" for "unknown determiner") is added. Note that this analytical decision still leaves open the possibility of [m] representing either a specific but unknown determiner (e.g. *ma* or *ama*) or a filler syllable (i.e. an element appearing in the pre-nominal slot).

- Record the analytic difficulty and the reasons for your analysis on a separate tier. We are aware that including such a tier is additional work. However, given the indeterminacy in analyzing children's utterances, it will be necessary to inspect each instance that is turned up by a corpus search. In this context, such a 'notes' tier allows you to recreate your analytical decisions (and to reconsider them, if necessary). In the case of recurring issues, we advise to note them in a separate text document. For example, nasal elements in pre-nominal slots are a recurring issue in Qaqet child language. We first detected them because the interlinearization process forced us to make a decision, which we documented on a 'notes' tier under the corresponding utterances. As the prevalence of these forms became obvious, we documented our decisions in a separate document, where we outlined the pros and cons of alternative decisions. This freed up the 'notes' tier, as we no longer added elaborate reasoning, but minimalistically referred to the separate document. Later, a corpus search allowed us to re-inspect our decisions and to refine them, and eventually allowed us to propose an analysis in our acquisition sketch.

Note that analyzing form X as morpheme Y is not intended to make any assumption about the child's 'intention'. The intention is not necessarily clear, and a child may not necessarily have the same phonemic and morphemic representations as an adult. We simply use the categories of the adult language as an approximation, with the goal of making the corpus searchable, which in turn is a prerequisite for detecting patterns (which in turn may put us into a position to reconsider some of our initial analytic decisions). Analyzing [am] as *-iam* is not supposed to make the assumption that the child 'intended' to use *-iam*. Rather it is specifying the context (as *-iam*), thus making it possible to search the corpus for noun class suffixes and possibly detecting patterns. We might come across similar contexts suggesting that *-iam* is often realized [am], or we might find that there are no clear instances of *-am*. If the context is not clear, an analysis as 'unknown morpheme' is preferable to no analysis. In our example, analyzing [m] as a determiner makes it possible to search the corpus for all determiners and compare their realizations, again enabling us to detect patterns.

This approach also guides against the danger of over-interpreting realizations as specific 'errors'. For example, the [am] realization above could be analyzed as a phonetic 'error' (simplifying the diphthong *ia*), a semantic 'error' (choosing the wrong noun class suffix) and/or a pragmatic 'error' (choosing a noun class suffix inappropriate to humans). The point is that we do not know and cannot make a decision on the basis of the one example. However, again, if we compare a larger number of contexts (e.g. those containing noun class marking or containing determiners), we may detect relevant patterns.

Finally, example (1) illustrates yet another common phenomenon: the presence of rote-learned forms, where a young child produces a morphologically complex form without having analyzed its internal structure. The above structure is morphologically very complex and it is unlikely that a child of 1;11 has productive command over it. However, it is impossible to decide on the basis of a single utterance, and our morphological analysis is not intended to make any assumption about productivity. Yet it does allow us to search the corpus for, say, the presence of noun class suffixes or the number of morphemes per utterance. Doing so will probably reveal that the corpus contains very few utterances of two-year-old children that contain nouns with noun class suffixes or that consist of four morphemes. Inspecting these utterances may then reveal contextual factors that suggest that they are unanalyzed forms occurring, for example, in repetition contexts.

In addition to a morphemic annotation, you may consider annotating other phenomena of interest. Systematic annotation is very time-consuming, though – both the annotation itself and the development or adaptation of an existing annotation scheme. Therefore, we do not recommend annotating the entire corpus for all the possible phenomena. We do recommend that you let yourself be guided by your interests, however, and that you annotate some parts of the corpus for those phenomena that are of interest to you. We offer some pointers and example analyses in Part II.

4 Summary

We have presented Part I of a manual to be used as a guide for anyone interested in working across child language and language documentation. This includes those foraying into the area of child language and collecting data for the first time, child language specialists collecting field-based data on understudied languages, and researchers investigating potential typological differences. In Part I, we have focused on constructing a sketch corpus: the structure of the corpus (in Section 2), and issues of data processing (in Section 3). In Part II, we present a model for developing a child language acquisition sketch: a section-by-section guide that offers suggestions for analyzing and presenting child language and child-directed language material.

References

Bates, Elizabeth, Philip S. Dale & Donna Thal. 1995. Individual differences and their implications for theories of language development. In Paul Fletcher & Brian MacWhinney (eds.), *Handbook of child language*, 96–151. Malden, MA: Blackwell. https://doi.org/10.1111/b.9780631203124.1996.00005.x

Bavin, Edith L. & Letitia R. Naigles (eds.). 2015. *The Cambridge handbook of child language.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316095829

Behrens, Heike (ed.). 2008. *Corpora in language acquisition research: History, methods, perspectives.* Amsterdam: John Benjamins. https://doi.org/10.1075/tilar.6

Berman, Ruth A. & Dan I. Slobin (eds.). 1994. *Relating events in narrative: A crosslinguistic developmental study.* Hillsdale, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9780203773512

Bowern, Claire. 2015. *Linguistic fieldwork: A practical guide* (2nd edition). London: Palgrave Macmillan. https://doi.org/10.1057/9781137340801

Brown, Roger. 1973. *A first language: The early stages.* Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Casillas, Marisa, Penelope Brown & Stephen C. Levinson. 2020. Early language experience in a Tseltal Mayan village. *Child Development* 91(5). 1819–1835. https://doi.org/10.1111/cdev.13349

Casillas, Marisa & Alejandrina Cristia. 2019. A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology* 5(1). 24. https://doi.org/10.1525/collabra.209

Child Language Research and Revitalization Working Group. 2017. *Language documentation, revitalization, and reclamation: Supporting young learners and their communities.* Waltham, MA: EDC.

CLAN (Version 05-Jan-2021) [Computer software]. 2021. TalkBank. Retrieved from https://dali.talkbank.org/clan/

Clark, Eve V. 2016. *First language acquisition* (3rd edition). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316534175

Cychosz, Meg, Adriana Weisleder, Rachel R. Romeo, Melanie Soderstrom, Alejandrina Cristia, Camilla Scaff, Janet Y. Bang, Marisa Casillas, Hillary Ganek & Kaya der Barbaro. 2020. *Daylong Audio Recording of Children's Linguistic Environments (DARCLE) Ethics Repository.* Open Science Framework Project: https://osf.io/u3tfv/

De Cat, Cecile, Draško Kašćelan, Philippe Prevost, Ludovica Serratrice, Laurie Tuller & Sharon Unsworth. 2021. Delphi consensus survey on how to document bilingual experience. https://doi.org/10.31219/osf.io/ebh3c

Demuth, Katherine. 1996. Collecting spontaneous production data. In Dana McDaniel, Cecile McKee & Helen Smith Cairns (eds.), *Methods of assessing children's syntax*, 3–22. Cambridge, MA: The MIT Press.

Demuth, Katherine. 2021. Managing acquisition data for developing large Sesotho, English and French corpora for CHILDES. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The open handbook of linguistic data management.* Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12200.003.0036

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197730.31

Eisenbeiß, Sonja. 2005. Documenting child language. In Peter K. Austin (ed.), *Language documentation and description* (vol. 3), 106–140. London: SOAS.

Eisenbeiß, Sonja. 2010. Production methods in language acquisition research. In Elma Blom & Sharon Unsworth (eds.), *Experimental methods in language acquisition research*, 11–34. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.27.03eis

ELAN (Version 6.0) [Computer software]. 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Evans, Nicholas. 2010. *Dying words: Endangered languages and what they have to tell us.* Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781444310450

Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5). 429–448. https://doi.org/10.1017/S0140525X0999094X

Fenson, Larry, Philip S. Dale, J. Steven Reznick, Elizabeth Bates, Donna J. Thal, Stephen J. Pethick, Michael Tomasello, Carolyn B. Mervis & Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development* 59(5). https://doi.org/10.2307/1166093

FieldWorks (Version 9.0) [Computer software]. 2020. SIL International. Retrieved from https://software.sil.org/fieldworks/

Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of language documentation.* Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197730

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne & Nora C. England. 1992. Endangered languages. *Language* 68(1). 1–42. https://doi.org/10.2307/416368

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195. https://doi.org/10.1515/ling.1998.36.1.161

Himmelmann, Nikolaus P. 2006a. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197730.1

Himmelmann, Nikolaus P. 2006b. The challenges of segmenting spoken language. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 253–274. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197730.253

Hoff, Erika. 2014. *Language development* (5th edition). Belmont, CA: Wadsworth Cengage Learning.

Innes, Pamela & Erin Debenport (eds.). 2010. Ethical dimensions of language documentation. *Language and Communication* 30(3). 159-210.

Jansco, Anna, Steven Moran & Sabine Stoll. 2020. The ACQDIV corpus database and aggregation pipeline. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, 11–16 May, 156–165.

Kidd, Evan & Seamus Donnelly. 2020. Individual differences in first language acquisition. *Annual Review of Linguistics* 6. 319–340. https://doi.org/10.1146/annurev-linguistics-011619-030326

Kidd, Evan & Rowena Garcia. 2022. How diverse is child language acquisition research? *First Language*. https://doi.org/10.1177/01427237211066405

Lieven, Elena V. M. 1997. Variation in a crosslinguistic context. In Dan I. Slobin (ed.), *The crosslinguistic study of language acquisition* (vol. 5), 199–263. Mahwah, NJ: Lawrence Erlbaum.

Lieven, Elena V. M., Heike Behrens, Jennifer Speares & Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30(2). 333–370. https://doi.org/10.1017/S0305000903005592

Lieven, Elena V. M. & Sabine Stoll. 2009. Language. In Marc H. Bornstein (ed.), *The handbook of cross-cultural developmental science*, 134–160. New York: Psychology Press.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk* (3rd edition). Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, Brian. 2021. *The CHILDES Project: Tools for Analyzing Talk. Part 1: The CHAT Transcription Format*. https://doi.org/10.21415/3mhn-0z89

MacWhinney, Brian & Elizabeth Bates (eds.). 1989. *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.

MacWhinney, Brian & William O'Grady (eds.). 2015. *The handbook of language emergence*. Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781118346136

McDonnell, Bradley, Andrea L. Berez-Kroeker & Gary Holton (eds.). 2018. Reflections on Language Documentation: 20 Years after Himmelmann 1998. *Language Documentation and Conservation SP 15*. http://hdl.handle.net/10125/24800

Meakins, Felicity, Jennifer Green & Myfany Turpin. 2018. *Understanding linguistic fieldwork*. New York: Routledge. https://doi.org/10.4324/9780203701294

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 67–85. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197730.67

Norcliffe, Elisabeth, Alice C. Harris & T. Florian Jaeger. 2015. Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience* 30(9). 1009–1032. https://doi.org/10.1080/23273798.2015.1080373

Peters, Ann. 2001. Filler syllables: What is their status in emerging grammar? *Journal of Child Language* 28(1). 229–242. https://doi.org/10.1017/S0305000900004438

Pye, Clifton. 2017. *The comparative method of language acquisition research*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226481319.001.0001

Pye, Clifton. 2021. Documenting the acquisition of indigenous languages. *Journal of Child Language* 48 (3). 454–479. https://doi.org/10.1017/S0305000920000318

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4(1-4). 123–155. https://doi.org/10.1007/s10805-006-9016-2

Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation and Conservation* 5. 187–207. http://hdl.handle.net/10125/4498

Rice, Keren. 2012. Ethical issues in linguistic fieldwork. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 407–429. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571888.013.0019

Rowland, Caroline F. 2014. *Understanding child language acquisition*. Abingdon: Routledge. https://doi.org/10.4324/9780203776025

Rowland, Caroline F. & Sarah L. Fletcher. 2006. The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language* 33(4). 859–877. https://doi.org/10.1017/S0305000906007537

Simons, Gary F. & Charles D. Fennig (eds.). 2017. *Ethnologue: Languages of the world* (20th edition). Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Slobin, Dan I. 1985-1997. *The crosslinguistic study of language acquisition* (vol. 1–5). Mahwah, NJ: Lawrence Erlbaum.

Slobin, Dan I., Susan M. Ervin-Tripp, John J. Gumperz, Jan Brukman, Keith Kernan, Claudia Mitchell & Brian Stross. 1967. *A field manual for cross-cultural study of the acquisition of communicative competence* (second draft). Berkeley: University of California at Berkeley.

Thieberger, Nicholas (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571888.001.0001

Tomasello, Michael & Daniel Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language* 31(1). 101–121. https://doi.org/10.1017/S0305000903005944

Toolbox (Version 1.6.4) [Computer software]. 2019. SIL International. Retrieved from https://software.sil.org/toolbox/

Received: 12 Jul 2021
Accepted: 4 Oct 2022