Language Documentation & Conservation Special Publication 25
*Doing corpus-based typology with spoken language data: State of the art*
edited by Geoffrey Haig, Stefan Schnell, and Frank Seifart, pp. 29 – 58
https://nflrc.hawaii.edu/ldc/sp25
https://hdl.handle.net/10125/74657

# Child language documentation: The sketch acquisition project

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

[1] *University of Cologne,* [2] *University of Melbourne,*
[3] *Max Planck Institute for Psycholinguistics,* [4] *Australian National University,*
[5] *University of Kaiserslautern*

## Abstract

This paper reports on an on-going project designed to collect comparable corpus data on child language and child-directed language in under-researched languages. Despite a long history of cross-linguistic research, there is a severe empirical bias within language acquisition research: Data is available for less than 2% of the world's languages, heavily skewed towards the larger and better-described languages. As a result, theories of language development tend to be grounded in a non-representative sample, and we know little about the acquisition of typologically-diverse languages from different families, regions, or sociocultural contexts. It is very likely that the reasons are to be found in the forbidding methodological challenges of constructing child language corpora under fieldwork conditions with their strict requirements on participant selection, sampling intervals, and amounts of data. There is thus an urgent need for proposals that facilitate and encourage language acquisition research across a wide variety of languages. Adopting a language documentation perspective, we illustrate an approach that combines the construction of manageable corpora of natural interaction with and between children with a sketch description of the corpus data – resulting in a set of comparable corpora and comparable sketches that form the basis for cross-linguistic comparisons.

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

# 1 Introduction

This paper contributes a language acquisition perspective to the topic of this special issue on cross-corpus typologies. It introduces a project designed to collect comparable corpus data on child language and child-directed language in under-researched languages. Before discussing the project in Section 2, this introductory section gives background information and addresses the motivation behind the project.[1]

Corpora of natural language use are considered important data sources for language acquisition research (see, e.g., the CHILDES database; MacWhinney 2000). In particular, longitudinal corpora play a central role: Such corpora follow the same child(ren) over long periods and thereby trace the development of each child individually. At the same time, such corpora constitute valuable resources for more experimental approaches that systematically

[1] The sketch acquisition project was developed over a number of years. We sincerely thank the participants at a number of workshops and conferences for engaging with this project and for enriching it with their different perspectives and backgrounds, as well as the many colleagues who commented on various draft proposals of the sketch format. We gratefully acknowledge funding from the joint German Academic Exchange Service and Universities Australia funding scheme (2019–2021, *A guide for documenting language development in under-studied languages*, grant no. 57446167). In Papua New Guinea, our thanks go to the Qaqet communities of Raunsepna and Lamarain for their generosity and help, and to the Volkswagen Foundation's Lichtenberg program (Az 87 100), the Endangered Languages Documentation Programme (SG0110) and the Australian Research Council (FT0991412) for their generous funding. Many thanks to Geoffrey Haig, Pavel Ozerov, Stefan Schnell, Frank Seifart, and an anonymous reviewer for critical feedback on this paper.

probe children's knowledge of specific aspects of their language (Blom & Unsworth 2010; Eisenbeiß 2006, 2010). The importance of corpora is reflected in a large body of literature on principles of corpus design (e.g. Behrens 2008; Demuth 1996, 2008; Eisenbeiß 2006, 2010; Parisse 2019; Tomasello & Stahl 2004).

This literature addresses the specific challenges of child language, and we highlight two central issues here. First, issues relating to the participants: the number of participating children, their age, developmental stage, and gender, amongst many other criteria for a child's in- or exclusion in a sample. This is of importance as there are considerable differences in individual children's development, impacting our ability to generalize from one child to other children of the same age. And second, issues relating to the overall amount of data and sampling intervals. Ideally, acquisition corpora capture the child's developing knowledge, including not only the productive use of a form, but especially the developmental trajectory towards it. The interpretation of such data is not straightforward, though, as the presence of a form does not necessarily mean that the children know it and use it productively: They may have learned it as an unanalyzed form, and we need data on, for instance, systematic errors to guide our interpretation. Conversely, its absence does not necessarily mean that they do not know it: It may be a gap in the data. Sample size and sample density crucially influence the likelihood that a form appears in the data, or that systematic errors can be detected. A further line of research revolves around how distributions in the input influence acquisition, thus similarly necessitating large and balanced corpora.

Proposals differ in how they address the above issues, but the overall consensus is that large amounts of data are needed. For example, a current proposal (implemented in the Chintang corpus; see Stoll et al. 2009) makes a good case for 4–5 hours of recording each month, done over a short time span (e.g. in the first week of each month) continuously over a longer period, and including more than one focus child. Adhering to such a schedule, and recording two children for two years, would result in 192–240 hours of recording – an amount of data that is likely to be beyond the capabilities of even the most ambitious language documentation project. And, as Tomasello & Stahl (2004: 118)) remark, even this amount of data may not be enough for some research questions: "[…] [T]he majority of existing child speech samples […] represent only a very small proportion of all the language the child produces and hears – on average around 1%. [...] [A]nd in some cases 1% sampling is not adequate to answer the question at hand."

Against this background, it is not surprising that research is heavily biased towards better-researched languages. It is generally estimated that acquisition studies are available for 1–2% of the world's languages; and these estimates include languages for which only one article is published (see Kelly & Nordlinger 2014; Lieven & Stoll 2009; Slobin 2014; Stoll & Bickel 2013; for psycholinguistic research in general, see Anand et al. 2010). Recently, Kidd & Garcia (forthcoming) systematically surveyed studies published until 2020 in the four major acquisition journals, and they find that contributions cover 1.47–1.72% of the world's languages – with the majority on English (54%). Other Indo-European languages account for 30% (skewed towards large Romance and Germanic languages), and non-Indo-European languages for the remaining 16% (skewed towards languages with a large speaker and researcher base, such as Hebrew, Mandarin, Japanese).

The available data sources thus do not reflect the enormous diversity of our world with its 7000+ languages. And this in turn impacts our ability to address one of the central challenges to linguistics, which Evans & Levinson (2009: 447) put as follows: "[T]o show how the child's mind can learn and the adult's mind can use, with approximately equal ease, any one of this vast range of alternative systems. [...] [This] calls for a diversified and strategic harnessing of linguistic diversity as the independent variable in studying language acquisition and language processing [...]: Can different systems be acquired by the same learning strategies, are learning rates really equivalent, and are some types of structure in fact easier to use?"

A number of initiatives have addressed this empirical bias over the years, such as Dan Slobin's efforts in the *Crosslinguistic study of language acquisition* series (Slobin 1985a–1997b) and the Frog Story project (Berman & Slobin 1994; Strömqvist & Verhoeven 2004), Brian MacWhinney's CHILDES project (MacWhinney 2000), and Bambi Schieffelin's and Elinor Ochs' language socialization paradigm (Schieffelin & Ochs 1986). Slobin & Bowerman (2007), Bowerman (2011), and Berman (2014) give succinct summaries of cross-linguistic acquisition research conducted within the above traditions, showing the important contribution of data and analyses from typologically-diverse languages for theories of language development. But they also acknowledge that "collecting and processing primary data is extremely expensive in terms of time, money, and personnel, so researchers nowadays often work with published or online transcripts of child-caregiver interaction [...], with comparisons limited to a small set of languages" (Slobin & Bowerman 2007: 216). Since the publication of the above contributions, further child language corpora of diverse languages were compiled, or are in the process of being compiled, and are being utilized in cross-corpus research, most notably within Sabine Stoll's ACQDIV project at the University of Zurich (Jansco et al. 2020; Stoll & Bickel 2013).

It is beyond doubt that there is an urgent need for child language corpora of the above type: corpora of diverse languages that adhere to the strict requirements for the construction of acquisition corpora. But it is similarly clear that these requirements will cause insurmountable difficulties in many fieldsites and that the number of such corpora will continue to remain small. We will thus have to devise supplementary approaches for diversifying the evidential base of language acquisition research. Recent years have seen two very promising proposals that are set to encourage acquisition research on under-researched languages: a focus on day-long recordings (Casillas & Cristia 2019) and a toolkit structured around basic linguistic phenomena (Pye 2021). In the next section, we propose a further supplementary approach, and report on an on-going project that explores the possibility of creating small-scale child language corpora. We introduce the corpus design (Section 2.1) and illustrate its potential by means of a case study (Section 2.2) before summarizing the discussion in Section 3.

Given the efforts involved in constructing corpora of under-researched languages, we envision such corpora to serve multiple purposes: to feed into theories of language development, but also to contribute to the documentation of languages and to support communities in language maintenance and revitalization projects. The focus of this paper is on the first issue. For a more general documentation perspective, see Hellwig & Jung (2020); and for a maintenance and revitalization perspective, see the Child Language Research and Revitalization Working Group (2017).

## 2 The sketch acquisition project

This section introduces the sketch acquisition project. The project takes its inspiration from a central idea proposed by Dan Slobin and colleagues in their 1967 *Field manual for cross-cultural study of the acquisition of communicative competence*: "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). Slobin and colleagues' manual has a fairly broad scope. Its target audience are investigators conducting dedicated acquisition and socialization research, and they propose a 12-month fieldwork schedule for collecting different types of data, including spontaneous data. Their manual forms the basis for the highly successful *Crosslinguistic study of language acquisition* series (Slobin 1985a–1997b). Our own proposal is more limited in scope, and our target audience are language documenters and language communities who are interested in including a child language component in their documentations. At the same time, our proposal goes beyond Slobin and colleagues' original idea, as it draws on the advances and experience of language documentation over the last twenty years, adopting a language documentation perspective on data collection and focusing on the construction of comparable corpora of spontaneous language.

The sketch format was developed over the course of several digital and analogue meetings, including two workshops on the acquisition of lesser-documented languages at the Universities of Cologne (January 2019) and Melbourne (August 2019). We are currently piloting this approach in a number of languages, and we are preparing a special publication, including a sketch manual as well as examples of sketch descriptions and sketch corpora (Defina et al., forthcoming). The format consists of specific recommendations for corpus construction (e.g. identifying participants, amounts of data to be recorded and recording intervals, possible software and data formats), data processing (e.g. issues that arise when transcribing, translating, and annotating child language), and description (of different topics in child language and child-directed language, and for different audiences). The key component of our approach is flexibility, that is, we do not aim for a collection of corpora that are constructed and processed in identical ways. Our overall approach is similar to that of language documentation: a set of common principles and guidelines, but with flexibility in how corpus compilers implement them. We thus strive for a balance between comparability (through the guidelines) and flexibility (to cater for different eventualities as well as for different backgrounds, interests and motivations of researchers and communities). Flexibility, of course, interferes with any automatized comparison across corpora. However, we believe that a balance between comparability and flexibility is essential if we want to facilitate and encourage the collection of child data across a wide range of languages.

This paper introduces the sketch acquisition project and summarizes its main cornerstones. It revolves around compiling comparable child corpora of five hours of annotated spontaneous language (introduced in Section 2.1), combined with a sketch description of child language and child-directed language attested in this corpus (illustrated in Section 2.2). Throughout this paper, we exemplify our approach with sketch data from Qaqet [ISO 639-3: byx; glottocode qaqe1238]. Qaqet is spoken by 15 000 people in Papua New Guinea's East New Britain Province. It belongs to the geographically defined group of East Papuan languages (i.e. the approximately 25 non-Austronesian languages of Island Melanesia; Dunn et al. 2002), and it is part of the Baining language family (Stebbins 2009).

## 2.1 Overview

The sketch acquisition project builds on a sketch corpus that, ideally, approximates a longitudinal scenario (summarized in Table 1). This section discusses its implementation, and introduces the Qaqet sketch corpus and the amount of data available for analysis. To be of manageable size, the sketch corpus is limited to five hours of analyzed video-recorded spontaneous language (60 minutes at five different ages). The analyzed data is taken from a larger amount of recorded data. By recording more than five hours, we increase our chances that there is enough quality data available for analysis, and that we can exclude parts with little language. We do not envision a limit to the amount of recorded data. It should be minimally ten hours (120 minutes at five different ages), but ideally more, as this data constitutes a

| 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) |

**Table 1** Sketch corpus – longitudinal scenario. 30(60) stands for 30 minutes of analyzed data out of 60 minutes of recorded data.

valuable resource in its own right. The focus is on children aged 2;0[2] to 4;0, that is, on the period when large parts of a language are being acquired. It is known that individual children differ considerably, and that age can only serve as a rough indicator of development. We therefore approximate a longitudinal setup. Ideally, the same child is recorded at six-month intervals, so that it is possible to compare across different ages. To enable some limited form of comparability across individuals, the corpus contains data from two focus children.

In a given fieldwork context, it may not be feasible to follow the above scenario, though, and the template allows for flexibility. The most important factor is to identify families who are happy and willing to participate. The template therefore does not introduce any further requirements, for example, the children can be of any gender, and they need not be monolingual. They should, however, be identified by the community as not showing any signs of atypical development. And we strongly recommend giving preference to talkative children, as this maximizes the usefulness of the small data set. The children can be outside the envisioned target ages (two months older or younger), including children whose precise ages are not known. Furthermore, it may not be possible to record every six months. In this case, a cross-lagged approach is possible, that is, to record three or more children at two or three ages (e.g. to record the younger ages during a first fieldtrip, and then record the same children one year later during a second trip). If this approach

[2] Ages of children are given in the format YEAR;MONTH (e.g. 2;0 means an age of 2 years and 0 months).

| age (±2 months) | 2;0 | 2;6 | 3;0 | 3;6 | 4;0 |
|---|---|---|---|---|---|
| ZDL (male) | 2;1 | 2;8 | — | — | — |
| YDS (female) | 2;0 | 2;6 | — | — | — |
| YJL (female) | — | — | 3;1 | 3;7 | 4;0 |
| YRA (male) | — | — | 3;2 | 3;7 | 4;0 |
| data amount | 60(548) | 60(321) | 60(457) | 60(380) | 60(380) |

**Table 2** The Qaqet sketch corpus.

is not feasible either, a cross-sectional approach can be adopted, that is, to record ten different children at five different ages during the same trip.

Table 2 shows the implementation of the above template in Qaqet. It includes data from two younger and two older children, with some deviations from the target ages. For each child and age, 30 minutes of analyzed data are included. The number in brackets reflects the total amount of recorded data available for the two children (in roughly equal proportions for each child). The sketch data constitutes a subset from a larger longitudinal corpus that is still under construction, and it was selected to illustrate the cross-lagged approach, as we assume that this approach is very likely the most feasible approach in many fieldsites. The discussion in this paper is based on the sketch corpus, but it will sometimes make reference to the longitudinal corpus in order to situate observations in a larger context and to evaluate the usefulness of the sketch corpus.

All data was video-recorded by the parents. The aim was to capture natural interaction with the focus children, and it was left to the parents to decide what and when to record. The parents recorded predominantly in two types of settings. In one setting, recording took place outside of the family's house in the village. Here, the children were usually engaged in fairly stationary activities (playing with cards, stones, and toys; singing and dancing; eating; doing little household chores), and different adults and children tended to pass by and stop for a chat. In the other setting, recording took place in gardens outside of the village where families live semi-permanently. The children were usually engaged in more mobile activities (moving around the

| age (±2 months) | 2;0 | 2;6 | 3;0 | 3;6 | 4;0 | totals |
|---|---|---|---|---|---|---|
| IUs (focus children) | 448 | 510 | 605 | 331 | 711 | 2605 |
| IUs (other children) | 1101 | 601 | 476 | 665 | 612 | 3455 |
| IUs (adults) | 655 | 229 | 499 | 149 | 139 | 1671 |
| IUs (total) | 2204 | 1340 | 1580 | 1145 | 1462 | 7731 |
| words (total) | 3856 | 2277 | 2931 | 2271 | 2540 | 13875 |

**Table 3** Intonation units (IUs) in the Qaqet sketch corpus.

garden and digging for insects or vegetables; collecting wood and tending the fire; collecting and playing with sticks or feathers), and there was only ever a small number of people present. Anthropological observations suggest that these are two very typical settings in the Qaqet community, and both feature in the sketch corpus in roughly equal proportions, albeit unequally distributed across the different ages and children. A third typical setting, by contrast, is not represented: groups of children roaming on their own through the bush.

The sketch corpus thus gives insights into the children's learning environments. It features natural language, and it captures two of the three typical settings in which children grow up. But it has to be kept in mind that it is an opportunistic sample. There was no attempt to build a representative sample, or to control for factors such as number and type of interlocutors, kinds of activity, or location. Given the non-representativeness of the sample and the small amount of available data, rigorous quantitative approaches are problematic. Descriptive statistics are possible and desirable, though, as they show the distribution of phenomena in the corpus, provided that the numbers are interpreted with caution and are supplemented by qualitative analyses. Section 2.2 will illustrate some such analyses, and Section 3 will come back to the question of representativeness and sample size.

Table 3 gives an overview of the number of intonation units (IUs) and their distribution across the different participants and the five ages; it also includes an overall word count in order to give an idea of the available amount of language. The absolute and relative numbers vary considerably across the different ages, and also across the individual focus children (not captured in this table). Despite these differences, the table indicates a pattern that is also borne out by anthropological observations and by the larger corpus: Children spend much time with their peers, which is reflected in a larger number of intonation units uttered by other children (3 455, 44.7% of all units), compared to a smaller number produced by adults (1 671, 21.6%).

## 2.2 Sketch format

This section now turns to the acquisition sketch itself: the analysis of the five hours of sketch data. It discusses aspects of the language used by adults and older children when speaking with the focus children (Section 2.2.1) and of the language used by the focus children themselves (Section 2.2.2). Each discussion is structured around selected phenomena, and care was taken to cover a range of different phenomena from different levels of language. However, the selection remains arbitrary, and different phenomena could have been selected instead. The purpose of this section is not to be exhaustive, but to illustrate the potential of the sketch format by means of examples. All examples are taken from the Qaqet sketch corpus introduced in Tables 2 and 3. Wherever possible, descriptive statistics are used to show the distribution of a phenomenon in the sketch corpus and to convey an idea of how many tokens were available for analysis.

### 2.2.1 Child-directed language

When interacting with young children, adults are generally known to use a special register, that of child-directed language. This register is typically characterized by short (but correct and complete) utterances, few hesitations, exaggerated pitch contours, a high F0, long duration and pauses, a restricted vocabulary (and possibly a nursery vocabulary), a preference for questions and imperatives, and various forms of repetitions (e.g. Gallaway & Richards 1994; Snow & Ferguson 1977). But while the existence of this register and its typical features is well established, we still know little about its universality (Lieven 1994). It is known, though, that there are considerable differences in the relative proportion of overheard versus child-directed language and in the role of peer interaction (e.g. Cristia et al. 2019 for Tsimane; or Casillas et al.

| | adult | | older child | | totals | |
|---|---|---|---|---|---|---|
| varied repetition | 469 | 34.7% | 914 | 36.4% | 1 383 | 35.8% |
| exact repetition | 71 | 5.2% | 423 | 16.8% | 494 | 12.8% |
| total | 1 353 | | 2 511 | | 3 864 | |

**Table 4** Repetition and variation in child-directed intonation units.

2020 and 2021 for Tseltal and Yélî Dnye), and we know of counter-evidence to features proposed to be characteristic for child-directed language (e.g. Pye et al. 2017 report a low F0 for K'iche' Mayan). But the number of such studies is small, and the extent of variation remains unknown.

In contemporary acquisition research, the focus is on investigating if and how child-directed language and its features aid language development, with studies investigating correlations between input and output. While our sketch corpora are too small to allow for such analyses, they are big enough to contribute to our understanding of the variation space. Table 4 summarizes the distribution of child-directed language in the Qaqet sketch corpus: The corpus contains 3 864 such intonation units, that is, it contains enough data to allow for a qualitative analysis. This includes intonation units by adults (1 353, 35%) and by older children up to 6 years of age (2 511, 65%).[3] The proportions correspond to those attested in our larger longitudinal corpus, where two thirds of all child-directed intonation units originated in older children. The table also gives information on repetitions that will become relevant for the discussion later in this section.

The sketch format is unlikely to lend itself to a direct comparison of child- and adult-directed language, as there will be too little adult-to-adult language available for analysis, and as the contexts will not always be comparable. However, the sketch corpora are often set within larger documentation projects. In the case of Qaqet, a separate adult corpus exists and our knowledge of the adult language is advanced (Hellwig 2019). That is, although it is not

[3] The numbers only include intonation units directed from older to younger children, not those directed from younger to older children

possible to systematically compare the two registers, it is possible to describe the language used with children of different ages, to identify any conspicuous features and to form hypotheses about the existence of such a register and its differences to the adult language.

Against this background, this section now illustrates some of the possibilities of the sketch corpus, touching upon the following phenomena: a conversational routine, some prosodic features, and the structure of repetitions. An example of child-directed language is given in (1). In this example, a mother and her children are cleaning vegetables. Prior to this extract, the 2-year-old daughter is upset, and her mother now tries to comfort her. She does so by distracting her and talking about something that is of interest to the child. Such attempts at distraction are fairly common, and they often involve either going to the garden as in (1) or drawing the child's attention to some feature of the environment.[4]

**(1)** Mother and YDS (2;0):

**a.** **Mother:** *YDS, nautit*

| | |
|---|---|
| *YDS,* | *nani=ut=tit* |
| PN | can=1PL.SBJ=go.CONT |

'YDS, we can go'

**b.** *nani utit savramahleng*

| | | |
|---|---|---|
| *nani* | *ut=tit* | *se=pet=ama=sleng* |
| can | 1PL.SBJ=go.CONT | to=on/under=ART=garden |

'we can go to the garden'

[4] Morphological glossing follows the *Leipzig Glossing Rules*. Abbreviations: ART – article; CONT – continuous (aspect); DEM – demonstrative; DIST – distal; DU – dual; INTRG – interrogative; M – masculine; N – neuter; NCONT – non-continuous (aspect); NM – noun marker; NPST – non-past (tense); PL – plural; PN – proper name; POSS – possessive; PROX – proximal; PURP – purposive; SBJ – subject; SG – singular; SIM – simultaneous (conjunction).

c. *kua nyinarli?*

| *kua* | *nyi=narli* |
|---|---|
| INTRG | 2SG.SBJ.NPST=hear |

‘do you hear?’

d. *nautit sevanu savramahleng*

| *nani=ut=tit* | *se=panu* | *se=pet=ama=sleng* |
|---|---|---|
| can=1PL.SBJ=go.CONT | to=up | to=on/under=ART=garden |

‘we can go up to the garden’

e. *nani utit savramahleng*

| *nani* | *ut=tit* | *se=pet=ama=sleng* |
|---|---|---|
| can | 1PL.SBJ=go.CONT | to=on/under=ART=garden |

‘we can go to the garden’

f. **YDS:** *da?*

‘really?’

g. **Mother:** *ee*

‘yes’

h. *utir iv uretmatna*

| *ut=tit* | *ip* | *ure=tmatna* |
|---|---|---|
| 1PL.SBJ=go.CONT | PURP | 1PL.SBJ.NPST=do.work.CONT |

‘we go and work’

i. **YDS:** *da?*

‘really?’

j. **Mother:** *ee*

‘yes’ (LongYDS20150506_1 1161.956–1178.350)

Aside from the distraction routine, this little extract exemplifies a further routine: the tag question routine. Here, a child signals her participation in the conversation by asking the tag question *da?* ‘really?’ in (1f) and (1i), and the interlocutor responds with *ee* 'yes' in (1g) and (1j). This routine seems to constitute an important word learning context, as exemplified in (2). The older brother introduces a new word (*aqulavaska* 'kind of taro'), the child responds with the tag question, the brother acknowledges it, and the child then attempts (more or less successfully) to pronounce the new word. In the larger longitudinal corpus, the tag question is one of the most frequent utterances among children aged 2;0 to 2;2: It comes in fourth place, after the interjections *ah?* 'huh?', *mh ~ ee* 'yes (backchanneling)', and *ih!* 'surprise'. The sketch corpus accordingly contains many examples of this routine, making it possible to identify it and describe its distribution.

**(2)** Brother and YDS (2;6):

**a.** **Brother:** *giavaqaira amala.. amaqulavaska*

| | | |
|---|---|---|
| *gia=va-ka=iara* | *ama=la..* | *ama=qulavas-ka* |
| 2SG.POSS=thingy-SG.M=PROX | ART=?? | ART=taro-SG.M |

'your.. *qulavas* taro here'

**b.** **YDS:** *da?*

'really?'

**c.** **Brother:** *ee*

'yes'

**d.** **YDS:** *aqulaquis ka*

**e.** *avu.. avullai ka*

**f.** *aquista*

**g.** *avulaquistka*

**h.** *alu qavas*

'*qulavas* taro' (LongYDS20151204_1 483.705–495.365)

Example (1) exhibits further recurring features of child-directed language, including an exaggerated prosody. For example, Figure 1 gives the pitch contour for the polar question in (1c): It starts at a high frequency of 220 Hz, rises

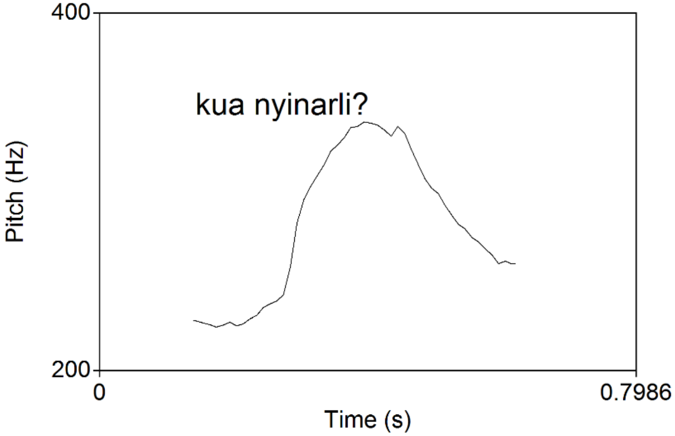

**Figure 1** Pitch contour for example (1c), *kua nyinarli?* 'do you hear?'

to 340 Hz, and falls again to 250 Hz. This rise/fall pitch contour is typical for polar questions in Qaqet (Hellwig 2019: 52–63), but the distinctions tend to be far less pronounced in adult-directed speech.

More generally, it is not unusual to encounter a careful pronunciation in the child-directed data. This can especially be observed across subsequent utterances, usually when a child is not responding as expected. For example, the two segmentally identical utterances in (1b) and (1e) (*nani utit savramahleng* 'we can go to the garden') are not produced identically: The second one is longer (1.11ms vs. 0.89ms), and the last syllable (*hleng* 'garden') is stressed. Compare also the contracted form *nautit* 'we can go' in lines (1a) and (1d) with its more careful pronunciation *nani utit* in lines (1b) and (1e) respectively.

Finally, example (1) features varied repetitions built around the verb *utit* 'we go'. The mother repeats this verb five times, but varies its context (adding or deleting a particle *nani* 'can', a directional *sevanu* 'up', a prepositional phrase *savramahleng* 'to the garden', and a purpose clause *iv uretmatna* 'and work'). This phenomenon is known in the literature as variation sets, that is, "partial repetitions [...], with changes in lexical items, grammatical morphology, and/or word order, maintaining a constant communicative intent" (Küntay & Slobin 1996: 267). Aside from varied repetitions, we also find exact repetitions in the corpus, as in (3). Here, the older brother produces the same utterance five times in a row.

**(3)** Brother to YDS (2;0):
*ulunyi* (5)
*ngu=lu-nyi*
1SG.SBJ.NPST=see.NCONT-2SG
'I see you (5x)' (LongYDS20150506_1 1352.975–1357.988)

As shown in Table 4, above, the sketch corpus contains many varied and exact repetitions, allowing for a description of the patterns: Who utters what to whom? In which contexts and with which functions? Which kinds of elements tend to be added, deleted, replaced, or re-ordered? And so on. It is likely that many sketch corpora will afford investigations into this phenomenon, as repetitions and especially variation sets have been shown to be frequent in many languages (e.g. Küntay & Slobin 1996; Moran et al. 2019; Onnis et al. 2008; Slobin et al. 2010; Wirén et al. 2016). Qaqet adults show a similar pattern to that reported in the literature: 25% or more of all utterances directed to young children tend to be in the form of variation sets, while exact repetitions are reported to be less frequent (around 1%) (see especially Wirén et al. 2016). Older Qaqet children, by contrast, produce many exact repetitions. Given that peers are their most frequent interlocutors, a considerable part of a Qaqet child's input thus consists of exact repetitions.

As said above, these numbers have to be interpreted with caution: They could simply be an artefact of the small corpus size. If borne out in a larger corpus,[5] such differences would be of interest to theories of learning.

[5] Indeed, the proportion of exact repetitions in our larger longitudinal corpus is similar (10% of child-directed intonation units, mainly uttered by older children). The proportion of varied repetitions is even lower in the longitudinal corpus (29% for adults, 16% for older children). The difference is likely to be due to one specific recording in the sketch corpus, which contains an exceptionally large amount of one-on-one interaction where an older child elicits body part terms from a younger child. It is likely that this kind of context favors varied repetitions.

Variation sets are known to facilitate language learning: Keeping the message constant, but varying its form, allows the learner to compare across subsequent utterances, identifying what remains identical and what changes. Exact repetitions, by contrast, may help segmentation if a known morpheme is part of the repetition, and they probably facilitate the storing of unanalyzed units. An unexpectedly large proportion of exact repetitions in the input thus raises questions about their effect on learning (e.g. different forms of repetition may be useful at different stages of learning), and it speaks to current debates about the amount and kind of input that facilitate the acquisition of a language. To be clear, the sketch corpus cannot answer these questions. But it is possible to describe the attested patterns, to formulate the hypothesis that Qaqet children interact preferably with other children and hear a larger proportion of exact repetitions than their Western peers, and to raise the question of what this pattern would mean for a theory of learning.

The sketch corpus is composed in such a way that it contains as many child-directed and children's utterances as possible. This approach has the advantage that it generates a large amount of analyzable data. But it also has its limitations. In particular, it biases the selected contexts towards child-directed interactions, and this is also true for the Qaqet sketch corpus. By contrast, it does not feature many multi-party interactions where children are present during interactions among adults. Our anthropological observations suggest that such contexts are not infrequent, and our larger longitudinal corpus shows that adults tend to not attend to children in these contexts, or preferably only interact with the oldest child present. That is, children are exposed to non-child-directed language to varying degrees, but these contexts are marginal in the sketch corpus.

Having added these qualifications, the data offers an excellent insight into how adults and older children talk to young children, making it possible to identify and describe numerous features of child-directed language. The goal of this section was to show what is possible on the basis of only five hours of data, and to thus exemplify the potential of the sketch data for a qualitative analysis of child-directed language. For Qaqet, there exists a controlled study that systematically compares child-directed and adult-directed language in re-tellings of the Pear Story (Frye 2019). We are thus in a position to evaluate the findings of our sketch corpus against the results of that study and all of the features identified in the systematic study are present and common in the sketch corpus. Going beyond Qaqet, it will eventually be possible to compare features of child-directed language across languages as more such corpora become available. We conducted a preliminary comparison across the sketch corpora of four languages of the Australia-Pacific region (Murrinhpatha, Pitjantjatjara, Nungon, Qaqet), showing the existence of this register, as well as similarities (e.g. in prosody and routines) and differences (e.g. in nursery vocabulary and morphosyntactic simplification) (Davidson et al. 2021).

### 2.2.2 Child language

Turning to child language now, examples (4) and (5) contrast typical 'give'-utterances from the youngest and oldest age groups.

**(4)** YDS (2;0):
*na*
hither
'give it [wants to be given a sweet potato]'
(LongYDS20150506_1 188.905–189.529)

**(5)** YJL (4;0):
*lira nguaqurlanyi tlungera*

| | | |
|---|---|---|
| *lira* | *ngua=quarl-nyi* | *te=lu-nget-a* |
| just.now | 1SG.SBJ=present/shine.NCONT-2SG | PURP=DEM-N-DIST |
| (i.) | (ii.)=(iii.)-(iv.) | (v.)=(vi.) |

'just now I gave you those ones [talking about peanuts]'
(LongYJL20150218_1 2358.030–2360.050)

In the youngest age group, one-word utterances predominate, and children talk about the here and now. In the oldest age group, by contrast, children competently make use of adult-like structures to talk about past, present, and future events. In example (5), YJL produces:

1. particles (i.),[6] choosing from approximately 40 particles (conveying tense, aspect, modality, mood, polarity etc.), which are combinable

[6] The indices in the list cross-refer to the numbered structures in (5).

in fixed orders;

2. a verb (iii.). In this case, a semantically-general three-place predicate whose interpretation of 'give' is triggered by the form of the recipient and theme arguments. The recipient (iv.) is realized as a pronominal suffix on the verb. And the theme (vi.) is introduced by a preposition (v.) chosen from a set of 13 available prepositions;
3. a subject index (ii.) procliticized to the verb (iii.), expressing person, number, and tense; for instance, YJL chooses the neutral index *ngua* that contrasts with the non-past index *ngu*;
4. the appropriate aspectual stem (iii.). Depending on the conjugation class, a verb has up to five stems that differ in their initial consonant(s); for instance, YJL chooses non-continuous *quarl*, contrasting with continuous *kuarl*;
5. a distal demonstrative (vi.) that is inflected for neuter noun class, choosing from eight classes and three number categories.

Example (5) shows some of the complexities of Qaqet grammar that 4-year-olds in the sketch corpus are quite capable of producing. In particular, their utterances differ considerably from what we see in the 2-year-olds, where we observe a small vocabulary, single-word utterances, and little morphology, as in (4). The sketch corpus allows us to describe what the children do at the various ages, and the remainder of this section gives three examples: the production of rhotics, the production of articles, and the realization of arguments.

The first example comes from phonology. The sketch corpus will give us an idea of how children of different ages realize the various sounds of their language. For example, rhotics tend to be challenging (e.g. Solé 2002), and we observe such difficulties in the Qaqet sketch corpus, too. Qaqet distinguishes between the retroflex flap *rl* (/ɽ/) and the alveolar trill *r* (/r/, realized as [r] ~ [ɾ]), and younger children tend to substitute both sounds. Table 5 shows typical realizations of frequent words in the sketch corpus.

The sketch data not only features many non-target-like forms, but also allows us to observe how interlocutors react to them. In Qaqet, it is common for interlocutors to imitate and to make fun of such forms, sometimes also to model the correct form. Example (6) illustrates all three possibilities, as the older cousin tries to get a 2-year old to produce the alveolar trill.

| gloss | adult realization | child (2;0–2;6) realization |
|---|---|---|
| PN | *arum* | *alum* |
| taro | *taru* | *talu ~ taju* |
| taro | *(a)rim* | *(a)jim* |
| many | *buɻum* | *bulum* |

**Table 5** Realization of *rl* (/ɻ/) and *r* (/r/, realized as [r] ~ [ɾ]) in the sketch corpus.

(6) Older cousin and ZDL (2;1):

| | | | |
|---|---|---|---|
| a. | **Cousin:** | *ia, ar* | ‘sorry, *ar*’ |
| b. | **ZDL:** | *al*$^{ə}$ | ‘*al*$^{ə}$’ |
| c. | **Cousin:** | [laugh] *al*$^{ə}$ [imitates] | ‘*al*$^{ə}$’ |
| d. | | *ar* | ‘*ar*’ |
| e. | **ZDL:** | *al*$^{ə}$ | ‘*al*$^{ə}$’ |
| f. | **Cousin:** | [laugh] | |
| g. | **ZDL:** | [laugh] | |
| h. | **Cousin:** | *nyi ma, ar* | ‘you now like this, *ar*’ |
| i. | **ZDL:** | *al*$^{ə}$ | ‘*al*$^{ə}$’ |
| j. | **Cousin:** | [laugh] *nyi ma, ar* | ‘you now like this, *ar*’ |
| k. | **ZDL:** | *al*$^{ə}$ | ‘*al*$^{ə}$’ |
| l. | **Cousin:** | *ar* | ‘*ar*’ |

(LongZDL20160213_1 408.455–426.417)

Another example comes from morphology. Qaqet nouns are obligatorily preceded by an article or possessor index. Young children, by contrast,

| | ages 2;0–2;6 | ages 3;6–4;0 |
|---|---|---|
| zero | 72 | 20 |
| *a* ‘NM’ | 21 | 15 |
| POSS | 12 | 59 |
| other ART | 6 | 42 |

**Table 6** Tokens of pre-nominal articles and possessor indices in the sketch corpus.

usually produce only the bare noun, and Table 6 gives an indication of the prevalence of this phenomenon in the sketch corpus. On the basis of this limited data set, one could hypothesize a possible developmental trajectory: Children start out by producing bare nouns, then produce the phonologically simple and semantically empty noun marker *a*, with other articles coming in only later.

Again, it is possible to observe the reaction of interlocutors to non-target-like forms. A typical reaction is illustrated in (7). The child does not utter an article and produces a non-target-like realization of *kuukuk* ‘sweet potato’ in (7a) and (7c). The mother then models the correct pronunciation, but omits the article (7b). In the adult language, this omission would be ungrammatical. In child-directed language, by contrast, such forms are not infrequent. Furthermore, the omission of the article has morphophonological consequences: All Qaqet articles are vowel-final, triggering the lenition of noun-initial voiceless plosives, for example, *kuukuk* ‘sweet potato’ lenites to *quukuk* [ɣuukuk] in this environment.[7] But since the mother omits the article in (7b), she produces the underlying plosive *k*. The acquisition of lenition seems to be a difficult issue, and even older children do not always produce target-like forms. For example, in (7d), the older brother (3;2) adds the article, but does not lenite the plosive.

[7] The target realization is *aquukuka* (*a=kuukuk-ka* ‘NM=sweet.potato-SG.M’). The first *k* lenites to *q* when preceded by an article. The second *k* constitutes a lexical exception to the lenition rule: The Qaqet lexicon contains a small number of lexemes that unexpectedly feature voiceless plosives in intervocalic position. And the last *k* is underlyingly geminate and predictably does not lenite.

(7) Mother, older brother (3;2), and YDS (2;0):

a. **YDS:** ***kaakak*** ‘sweet potato’

b. **Mother:** *ee,* ***kuukuka*** ‘yes, sweet potato’

c. **YDS:** ***kaakak*** ‘sweet potato’

d. **Brother:** ***akuukuka*** ‘a sweet potato’

(LongYDS20150516_1 297.285–300.910)

The final example is concerned with argument realization. Subject arguments are obligatorily indexed as proclitics on the verb, and object arguments are either realized as unmarked arguments or as prepositional arguments, depending on the verb. In our sketch corpus, the youngest children tend to only produce verb stems, without any arguments. A typical example is given in (8). The child produces the verb *tit* ‘go’ without the obligatory subject index. In such cases, interlocutors frequently expand on the child’s utterance, modelling the correct form. Here, the mother supplies the appropriate subject index (plus the resulting assimilation in voicing of the following plosive).

(8) Mother and YDS (2;0):

a. **YDS:** *tit* ‘go’

b. *gaka* ‘my friend’

c. **Mother:** *mh?* ‘yes?’

d. **YDS:** *tit* ‘go’

e. **Mother:** ***un****dit*
***un****=tit*
1DU.SBJ=go.CONT
‘we go’ (LongYDS20150716_1 693.170–700.420)

Older children, by contrast, typically produce the subject index. In their case, the main challenge revolves around the object arguments. In particular, they tend not to produce the preposition of prepositional arguments. For example, YJL produces the subject index and the theme in (9), but not the target preposition *se* 'to' (added in square brackets). In the adult language, this preposition is needed to introduce the theme of *suam* 'steal'. Similar observations hold for the larger longitudinal corpus (Hellwig 2021).

(9) YJL (3;1) to baby brother:
*ngusuam* [***se***]*giavake*

| | |
|---|---|
| *ngu=suam* | [*se=*]*gia=va-ka* |
| 1SG.SBJ.NPST=steal | [to=]2SG.POSS=thingy-SG.M |

'I steal your thingy' (LongYJL20150218_1 1016.610–1017.950)

This section has presented a small selection of child language phenomena that can be observed in the sketch corpus, with the purpose of illustrating the richness of the data. Again, the same qualifications apply as in the case of child-directed language. But despite the obvious limitations, the data is rich enough to describe the language used by the children, to identify phenomena that seem to pose a challenge for the children, to make cautious inferences about the productivity of a form, and to formulate hypotheses about the developmental trajectory that would need to be tested against a larger data set. Again, it will eventually become possible to compare across the sketch corpora of several languages. A comparison involving Qaqet has not yet been attempted, but Allen & Jung (2021) have conducted a first comparison across two polysynthetic languages (Dëne Sųłıné, Inuktitut), showing how differences in morphological structure influence acquisition (e.g. the presence of errors and omissions, the timing of acquisition).

# 3 Conclusion

This contribution has given a brief outline of the sketch acquisition project. With only five hours of annotated data, the sketch corpus constitutes a limited database, and this, of course, impacts the kinds of generalizations that can be made. As shown throughout Section 2, such a corpus lends itself to qualitative analyses of what children and their interlocutors are doing and saying in which kinds of contexts, including descriptive statistics, albeit not rigorous quantitative analyses. Despite its limited scope, we find that the patterns and structures detected in the Qaqet sketch corpus fairly closely match those attested in a larger longitudinal corpus and in a controlled study on child-directed language. That is, the sketch corpus constitutes a valid resource that allows us to identify and describe salient features of child-directed and child language. It is not the kind of data that allows us to discover every phenomenon (especially low-frequency phenomena), to demonstrate differences between adult- and child-directed language, or to draw firm conclusions about the child's knowledge of specific forms or structures. To do so, we would not only need much larger and denser data sets, but also different types of data, including experimental data.

While recognizing the limits of this approach, we see clear advantages – not least because it focuses on a type of "observable linguistic behavior" (Himmelmann 1998: 166) that has so far played only a marginal role within language documentation. The sketch data and descriptions thereby constitute valuable records of child-directed and child language in a specific community, which in turn may feed into the production of educational material for language maintenance purposes. From the perspective of language acquisition research, sketch data and descriptions give insights into acquisition and socialization across a large number of languages, thus broadening our understanding of the problem space and allowing us to generate further hypotheses: Different types of languages and learning environments pose different challenges to the child, and yet children everywhere learn their languages. A realistic contribution of the sketch format is thus to map out variation – a variation that eventually needs to be taken into account for our theories of language development. In addition, given their parallel nature, it will be possible to compare sketches and sketch data across languages, or within a language family or an area. Preliminary results of such comparisons were presented at the 2021 conference of the International Association for the Study of Child Language (Defina et al. 2021). Finally, some of the sketches are likely to be developed into in-depth acquisition projects, as they allow to explore the feasibility of child language research in a particular context, or as they serve to identify phenomena of special interest to be investigated further.

## References

Allen, Shanley & Jung, Dagmar. 2021. *Comparing sketch acquisition studies in Dene and Inuktitut*. Paper presented at the Conference of the International Association for the Study of Child Language (IASCL), 15–23 July 2021.

Anand, Pranav & Chung, Sandra & Wagers, Matthew. 2010. *Widening the net: Challenges for gathering linguistic data in the digital age*. Response to *NSF SBE 2020: Future Research in the Social, Behavioral and Economic Sciences planning activity*. (`https://people.ucsc.edu/~schung/anandchungwagers.pdf`).

Behrens, Heike (ed.). 2008. *Corpora in language acquisition research: History, methods, perspectives*. Amsterdam: Benjamins.

Berman, Ruth A. 2014. Cross-linguistic comparisons in child language research. *Journal of Child Language* 41(S1). 26–37.

Berman, Ruth A. & Slobin, Dan I. 1994. *Relating events in narrative: A cross-linguistic developmental study*. Mahwah, NJ: Erlbaum.

Blom, Elma & Unsworth, Sharon (eds.). 2010. *Experimental methods in language acquisition research*. Amsterdam: Benjamins.

Bowerman, Melissa. 2011. Linguistic typology and first language acquisition. In Song, Jae Jung (ed.), *The Oxford handbook of linguistic typology*, 591–617. Oxford: Oxford University Press.

Casillas, Marisa & Brown, Penelope & Levinson, Stephen C. 2020. Early language experience in a Tseltal Mayan village. *Child Development* 91(5). 1819–1835.

Casillas, Marisa & Brown, Penelope & Levinson, Stephen C. 2021. Early language experience in a Papuan community. *Journal of Child Language* 48(4). 792–814.

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

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

Cristia, Alejandrina & Dupoux, Emmanuel & Gurven, Michael & Stieglitz, Jonathan. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development* 90(3). 759–773.

Davidson, Lucy & Defina, Rebecca & Forshaw, Bill & Hellwig, Birgit & Sarvasy, Hannah & Wighton, Wanyima. 2021. *A comparison of child-directed speech in four languages of the Australia-Pacific region*. Paper presented at the Conference of the International Association for the Study of Child Language (IASCL), 15–23 July 2021.

Defina, Rebecca & Allen, Shanley & Davidson, Lucy & Hellwig, Birgit & Kelly, Barbara F. & Kidd, Evan (eds.). Forthcoming. *The sketch acquisition manual (SAM)* (*Language Documentation & Conservation* special publication). Honolulu, HI: University of Hawai'i Press.

Defina, Rebecca & Allen, Shanley & Davidson, Lucy & Hellwig, Birgit & Kelly, Barbara F. & Kidd, Evan. 2021. Can we describe non-WEIRD language development with only 5 hours? *Paper presented at the Conference of the International Association for the Study of Child Language (IASCL), 15–23 July 2021.*

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

Demuth, Katherine. 2008. Exploiting corpora for language acquisition research. In Behrens, Heike (ed.), *Corpora in language acquisition research*: *Finding structure in data*, 199–205. Amsterdam: Benjamins.

Dunn, Michael & Reesink, Ger & Terrill, Angela. 2002. The East Papuan languages: A preliminary typological appraisal. *Oceanic Linguistics* 41(1). 28–62.

Eisenbeiß, Sonja. 2006. Documenting child language. In Austin, Peter K. (ed.), *Language documentation and description, volume 3*, 106–140. London: SOAS.

Eisenbeiß, Sonja. 2010. Production methods. In Blom, Elma & Unsworth, Sharon (eds.), *Experimental methods in language acquisition research*, 11–34. Amsterdam: Benjamins.

Evans, Nicholas & Levinson, Stephen C. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32. 429–492.

Frye, Henrike. 2019. *Child-directed speech in Qaqet*. Ph.D. dissertation, University of Cologne.

Gallaway, Clare & Richards, Brian J. (eds.). 1994. *Input and interaction in language acquisition*. Cambridge: Cambridge University Press.

Hellwig, Birgit. 2019. *A grammar of Qaqet*. Berlin: De Gruyter Mouton.

Hellwig, Birgit. 2021. Initial observations on complex predicates in Qaqet children's language. *First Language* 41(4). 406–429.

Hellwig, Birgit & Jung, Dagmar. 2020. Child-directed language – and how it informs the documentation and description of the adult language. *Language Documentation and Conservation* 14. 188–214.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(2). 161–195.

Jansco, Anna & Moran, Steven & Stoll, Sabine. 2020. The ACQDIV corpus database and aggregation pipeline. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'20), Marseille, France, 11–16 May 2020*. 156–165.

Kelly, Barbara & Nordlinger, Rachel. 2014. Fieldwork and first language acquisition. In Gawne, Lauren & Vaughan, Jill (eds.), *Selected papers from the 44th Conference of the Australian Linguistic Society, 2013*, 178–192. Melbourne: University of Melbourne.

Kidd, Evan & Garcia, Rowena. Forthcoming. How diverse is child language acquisition research? *First Language*.

Küntay, Aylın & Slobin, Dan I. 1996. Listening to a Turkish mother: Some puzzles for acquisition. In Slobin, Dan I. & Gerhardt, Julie & Kyratzis, Amy & Guo, Jiansheng (eds.), *Social interaction, social context, and language*: *Essays in honor of Susan Ervin-Tripp*, 265–286. Hillsdale, NJ: Erlbaum.

Lieven, Elena. 1994. Crosslinguistic and crosscultural aspects of language addressed to children. In Gallaway, Clare & Richards, Brian J. (eds.), *Input and interaction in language acquisition*, 56–73. Cambridge: Cambridge University Press.

Lieven, Elena & Stoll, Sabine. 2009. Language. In Bornstein, Marc H. (ed.), *The handbook of cross-cultural developmental science*, 543–555. Mahwah, NJ: Erlbaum.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analysing talk*. Mahwah, NJ: Erlbaum.

Moran, Steven & Lester, Nicholas A. & Gordon, Heath & Küntay, Aylin & Pfeiler, Barbara & Allen, Shanley & Stoll, Sabine. 2019. Variation sets in maximally diverse languages. In Brown, Megan M. & Dailey, Brady (eds.), *Proceedings of the 43rd annual Boston University Conference on Language Development*, 427–440. Somerville, MA: Cascadilla Press.

Onnis, Luca & Waterfall, Heidi R. & Edelman, Shimon. 2008. Learn locally, act globally: Learning language from variation set cues. *Cognition* 109. 423–430.

Parisse, Christophe. 2019. How large should a dense corpus be for reliable studies in early language acquisition? *CogniTextes* 19.

Pye, Clifton. 2021. Documenting the acquisition of indigenous languages. *Journal of Child Language* 48. 454–479.

Pye, Clifton & Pfeiler, Barbara & Mateo Pedro, Pedro. 2017. Mayan language acquisition. In Aissen, Judith & Zavala Maldonado, Roberto & England, Nora C. (eds.), *The Mayan languages*, 19–42. London: Routledge.

Schieffelin, Bambi & Ochs, Elinor (eds.). 1986. *Language socialization across cultures*. Cambridge: Cambridge University Press.

Slobin, Dan I. (ed.). 1985a. *The crosslinguistic study of language acquisition, Volume 1: The data*. Mahwah, NJ: Erlbaum.

Slobin, Dan I. (ed.). 1985b. *The crosslinguistic study of language acquisition, Volume 2: Theoretical issues*. Mahwah, NJ: Erlbaum.

Slobin, Dan I. (ed.). 1993. *The crosslinguistic study of language acquisition, Volume 3*. Mahwah, NJ: Erlbaum.

Slobin, Dan I. (ed.). 1997a. *The crosslinguistic study of language acquisition, Volume 4.* Mahwah, NJ: Erlbaum.

Slobin, Dan I. (ed.). 1997b. *The crosslinguistic study of language acquisition, Volume 5: Expanding the contexts.* Mahwah, NJ: Erlbaum.

Slobin, Dan I. 2014. Before the beginning: The development of tools of the trade. *Journal of Child Language* 41(S1). 1–17.

Slobin, Dan I. & Bowerman, Melissa. 2007. Interfaces between linguistic typology and child language research. *Linguistic Typology* 11(1). 213–226.

Slobin, Dan I. & Bowerman, Melissa & Brown, Penelope & Eisenbeiß, Sonja & Narasimhan, Bhuvana. 2010. Putting things in places: Developmental consequences of linguistic typology. In Bohnemeyer, Jürgen & Pederson, Eric (eds.), *Event represantation in language and cognition*, 134–165. New York: Cambridge University Press.

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

Snow, Catherine E. & Ferguson, Charles A. 1977. *Talking to children: Language input and acquisition.* Cambridge: Cambridge University Press.

Solé, Maria-Josep. 2002. Aerodynamic characteristics of trills and phonological patterning. *Journal of Phonetics* 30. 655–688.

Stebbins, Tonya N. 2009. The Papuan languages of the Eastern Bismarcks: Migration, origins and connections. In Evans, Beth (ed.), *Discovering history through language. Papers in honour of Malcolm Ross*, 223–243. Canberra: Pacific Linguistics.

Stoll, Sabine & Bickel, Balthasar. 2013. Capturing diversity in language acquisition research. In Bickel, Balthasar & Grenoble, Lenore A. & Peterson, David A. & Timberlake, Alan (eds.), *Language typology and historical contingency*, 195–216. Amsterdam: Benjamins.

Stoll, Sabine & Bickel, Balthasar & Lieven, Elena & Banjade, Goma & Bhatta, Toya Nath & Gaenszle, Martin & Paudyal, Netra Prasad & Rai, Manoj & Rai, Novel Kishore & Rai, Ichchha Purna. 2009. *Audiovisual Chintang corpus on language acquisition of 6 children. Electronic Database at the DoBeS archive.*

Strömqvist, Sven & Verhoeven, Ludo (eds.). 2004. *Relating events in narrative, Volume 2: Typological and contextual perspectives.* Mahwah, NJ: Erlbaum.

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

Wirén, Mats & Björkenstam, Kristina Nilsson & Grigonytė, Gintarė & Cortes, Elisabet Eir. 2016. Longitudinal studies of variation sets in child-directed speech. In Korhonen, Anna & Lenci, Alessandro & Murphy, Brian & Poibeau, Thierry & Villavicencio, Aline (eds.), *Proceedings of the 7th Workshop on Cognitive Aspects of*

*Computational Language Learning, Berlin, Germany, 11 August 2016*, 44–52. Berlin: Association for Computational Linguistics.