Efficient estimation of children's language exposure in two bilingual communities

Margaret Cychosz[1, 2], Anele Villanueva[3], and Adriana Weisleder[3]

[1]Department of Hearing and Speech Sciences, University of Maryland-College Park, College Park, MD

[2]Center for Comparative and Evolutionary Biology of Hearing, College Park, MD

[3]Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

Author Note

Author to whom correspondence should be addressed: Margaret Cychosz, 0100 Samuel J. LeFrak Hall, University of Maryland, College Park, MD, USA, 20742. Email: mcychosz@umd.edu

Abstract

Purpose: The language that children hear early in life predicts their later speech-language outcomes (Hoff 2003; Weisleder & Fernald 2013). This line of research relies on naturalistic observations of children's language input, often captured with daylong audio recordings. But the large quantity of data that daylong recordings generate requires novel analytical tools to feasibly parse thousands of hours of naturalistic speech. This study outlines a workflow to efficiently process and sample from daylong audio recordings made in two bilingual communities: Spanish-English in the United States and Quechua-Spanish in Bolivia.

Method: We employed a general sampling with replacement technique to efficiently estimate two key elements of children's early language environments: 1) proportion of child-directed speech and 2) dual language exposure. Proportions estimated from random sampling of 30-second segments were compared to those from annotations over the entire daylong recording (every-other-segment), as well as parental report.

Results: Results showed that approximately 49 minutes from each recording, or just 7% of the overall recording, were required to reach a stable proportion of child-directed speech and bilingual exposure. In both speech communities, strong correlations were found between bilingual language estimates made using random sampling and all-day annotation techniques. A strong relationship was additionally found for child-directed speech estimates in the United States, but this was weaker at the Bolivian site, where child-directed speech was less frequent. Furthermore, dual language estimates from the daylong audio recordings did not correspond to estimates derived from parental report.

Conclusions: Random sampling is a valid method to estimate ambient characteristics from daylong recordings. However, caution should be taken when interpreting estimates of low-frequency categories and practitioners might consider collecting multiple daylong recordings to accurately estimate characteristics of children's language exposure .

Keywords: bilingualism; child-directed speech; language acquisition; naturalistic recording; random sampling; LENA; Spanish

1 Introduction

In North America, the quantity and quality of language input that children receive can predict later language outcomes such as vocabulary size and lexical processing speed (Hart & Risley 1995; Hoff 2003; Ramírez-Esparza et al. 2014; Weisleder & Fernald 2013). In particular, studies have shown that the amount of child-directed speech (CDS) that children are exposed to during infancy and childhood has meaningful downstream effects on their language development (Hurtado et al. 2008; Mahr & Edwards 2018; Newman et al. 2016; Ramírez-Esparza et al. 2017; Rowe 2008; 2012; Weisleder & Fernald 2013). Similarly, differences in bilingual language exposure predict differences in infants' and children's speech processing, speech production, and vocabulary development, suggesting that the amount of bilingual children's exposure to each of their languages explains meaningful variability in their language development (Bijeljac-Babic et al. 2012; Byers-Heinlein 2013; Carbajal & Peperkamp 2019; Marchman et al. 2017; Orena et al. 2020; Pearson et al. 1997; Place & Hoff 2011; 2016; Potter et al. 2019; Unsworth et al. 2018; cf. Carroll 2017).

Traditionally, the methods for estimating children's bilingual and CDS exposure have differed. While children's bilingual language exposure has typically been estimated from parental report (Byers-Heinlein et al. 2019; DeAnda et al. 2016; de Houwer 2011; Place & Hoff 2011; 2016), CDS exposure has traditionally been measured from transcribed observations of parent-child interactions (Bergelson, Casillas, et al. 2019; Hirsh-Pasek et al. 2015; Hoff 2003). More recently, estimates of exposure to CDS and dual language input have been made using daylong audio recordings (8-16 hours) of children's everyday language environments (Marchman et al. 2017; Orena et al. 2019; Ramírez-Esparza et al. 2017) using the Language ENvironment Analysis system (LENA) (Gilkerson & Richards 2009). Yet while the LENA^{TM} system estimates the amount of adult speech near the child (e.g., the number of adult words), it does not differentiate between child- and adult-directed speech or between different languages. Previous studies have thus estimated the amount of CDS or dual language exposure in daylong recordings through human annotation. However, annotating the entirety of daylong recordings is time- and resource-intensive. This study proposes the use of random sampling of daylong recordings to more efficiently estimate children's speech-language exposure. We test this random sampling method in recordings collected from bilingual communities in the United States (Spanish-English) and Bolivia (Quechua-Spanish). This work first shows that stable proportions of CDS and bilingual exposure can be reached using random sampling after a modest amount of annotation. Then, the accuracy of estimates derived via random sampling is evaluated by comparing them to estimates made more traditionally via: 1) annotations of every-other-segment from the daylong recording and 2) bilingual exposure reported in parental language use questionnaires.

2 Background

2.1 Daylong recording methods

Recent years have seen an increase in studies that examine children's language experience and input based on large-scale, naturalistic observations. Many of these studies use daylong audio recordings, where children wear small, lightweight recorders over the course of an entire day, typically 8-16 hours.[1] These audio recordings, which capture what the child hears and says over the course of the day, can additionally be combined with other modalities such as video (Bergelson, Amatuni, et al. 2019) or photography (Casillas et al. 2019).

There are several advantages to quantifying elements of the child's environment with these daylong recordings. Perhaps most importantly, the derived analyses may be less prone to biases present in language questionnaires, such as recall or social desirability, or to observer effects during short observations (Bergelson, Amatuni, et al. 2019). The larger the naturalistic observation period, the more reliable and ecologically-valid a measurement of language use becomes, since it is difficult for participants to continuously self-monitor their behavior (Cychosz et al. 2020). Due to advances in technology, daylong recordings are also relatively easy to collect, which facilitates the involvement of understudied groups in behavioral research. This increased involvement of understudied groups is apparent in recent work employing daylong recordings in a wide variety of linguistic and cultural settings (Casillas et al. 2019; Casillas et al. 2020; Cristià et al. 2017; Cychosz 2020; Ganek et al. 2018; Orena et al. 2020; Ramírez-Esparza et al. 2017; Weisleder & Fernald 2013).

2.2 Estimating exposure to child-directed speech to predict learning outcomes

Because the quantity and quality of CDS in children's environments explain some individual variability in language-learning outcomes, an increasing number of studies are using daylong audio recordings to estimate CDS exposure. The quantity of CDS in children's environments can be estimated from daylong recordings 1) algorithmically, 2) manually via transcription and annotation, or 3) via a combination of these approaches.

A number of recent examples have algorithmically estimated children's exposure to adult speech. For example, Mahr & Edwards (2018) used estimates of the number of adult words per hour in the daylong recordings of children 2;4-3;3 derived from the LENA system's algorithms. LENA's algorithms distinguish between adult speech that is near and clear to the child (what they call "meaningful speech") and speech that is farther away or overlapping. Mahr & Edwards (2018) used LENA's counts of adult words in meaningful speech as an estimate of CDS, and showed that children with more adult words in their recordings had larger receptive vocabularies one year later. However, because the researchers did not manually annotate the recordings, they did not distinguish between child-directed and adult-directed speech near the child. Romeo et al. (2018) likewise used LENA's algorithms to automate CDS estimates, specifically the quantity of adult words and conversational turns between adults and the target child (aged 4;0-6;0). More conversational turns between adults and children—but not the total number of adult words—predicted increased language-related (Broca's area) neural activation during a language-listening task conducted in the lab. This result led the authors to suggest that LENA's automated conversational turn count better approximates CDS than the automated adult word count.

Other studies have used a combination of manual and algorithmic tools (Ramírez-Esparza et al. 2014; Weisleder & Fernald 2013; Marchman et al. 2017). For example, Weisleder & Fernald (2013) used LENA to process daylong recordings collected from Spanish-speaking infants at 1;7. Recordings were then divided into five-minute segments, which were manually classified as primarily containing child-directed or adult-directed speech. Then, the number of LENA-estimated adult words in each five-minute CDS segment was divided by the length of the recording to estimate each child's average hourly exposure to adult words in CDS. The authors found that more CDS at 1;7 predicted faster lexical processing and larger expressive vocabularies at 2;0. Similarly, Ramírez-Esparza et al. (2014) annotated 30-second segments from daylong recordings to indicate whether they contained infant-directed or adult-directed speech. They then used relative time-use estimates of infant-directed speech by calculating the proportion of intervals coded for each category. Results showed that infant-directed speech in one-on-one contexts was positively correlated with infant vocalizations and later vocabulary size.

Given the importance of CDS for learning outcomes in North America, researchers have also been interested in comparing the quantity of CDS from daylong recordings across distinct socio-cultural settings. Because LENA's algorithms were only trained on English, and its estimates have only been validated for a small number of languages, researchers doing cross-linguistic and cross-cultural work have turned to manual annotation and transcription. However, the time-intensive nature of manual transcription means that researchers often use automated estimates to guide decisions about which segments to transcribe. For example, one study used daylong recordings collected using LENA to estimate North American infants' (0;3-1;8) CDS exposure (Bergelson, Casillas, et al. 2019). First, 20 conversational blocks that LENA's algorithm identified as having at least 10 female and male adult speech tags near the child were selected from each recording. Each female or male speech tag was then manually annotated for speech register (child-directed or adult-directed). These annotations were used to measure the ratio of CDS to adult-directed speech over development, as well as the gender distribution of adult speech in the child's environment. Elsewhere, Casillas et al. (2019) employed only manual transcriptions to estimate CDS in recordings of children aged 0;2-3;0 in a Tseltal Mayan community in Mexico. Portions of each recording were selectively sampled and transcribed to obtain CDS quantities. Though CDS quantities were low—between one and five minutes per hour—the children met expressive language milestones, such as first word combinations, based on data from Western children. (See also Casillas et al. (2020) for similar work in a Yélî-Dnye community in Papua New Guinea and Cristià et al. (2017) for the Tsimane' in Bolivia.)

2.3 Dual language exposure predicts learning outcomes

Monolingual children contend with variability in the ambient speech stream from multiple speakers, who may modify their pitch and speaking rate from one production to the next. Bilingual children contend with this variation and more, since the speech stream contains two different languages (de Bree et al. 2017). Yet bilingual learners' experiences are far from uniform. Bilingual language environments vary widely with regards to the amount of input in each language, how the languages are separated by speaker or context, the amount of mixed input, and other properties of the ambient language (Byers-Heinlein 2013; Place & Hoff 2011). Importantly, many of bilingual children's learning outcomes—notably vocabulary —appear to depend upon the relative amount of exposure that the child receives in each of their languages. For example, Pearson et al. (1997) correlated the relative amount of Spanish-English input, measured via a language background questionnaire, with expressive vocabulary scores in children aged 0;8-2;6. They found strong correlations between vocabulary sizes in Spanish and English and the relative amount of input that children received in each language. In another study of Spanish-English bilinguals, caregivers of two-year-olds were asked to keep diary records of their children's language exposure (Place & Hoff 2011). Children who were reported to hear more Spanish had higher expressive vocabularies and greater grammatical complexity in Spanish, while children who were reported to hear more English performed better on the corresponding measures in English. (See also Place & Hoff (2016) who quantified language exposure via number of speakers and language mixing, and Byers-Heinlein (2013) who estimated the effect of language mixing on vocabulary development.)

Another study found bilingual exposure effects on the receptive vocabularies of bilingual French infants, aged 0;11 (Carbajal & Peperkamp 2019). Exposure was measured with a language questionnaire and two daily diaries to note who spoke in what language with the infant every half hour throughout the day. There was notable variability in the proportion of French in the environment, the frequency that French and the second language appeared in each half hour sequence, and the number of caregivers providing input. This variability was reflected in the infants' vocabularies: more maternal input in the second language resulted in a larger receptive vocabulary in that language. Using a detailed language use background questionnaire, Thordardottir (2011) also found exposure effects on receptive and expressive vocabulary in bilingual children aged 5;0 acquiring French and English. Children who were exposed to more French had larger expressive French vocabularies, and vice versa for English.

Similar exposure effects have been found for young children's speech processing and perception. In a looking-while-listening task, Potter et al. (2019) measured how bilingual English-Spanish children aged 1;6-2;6 processed words in their dominant and non-dominant languages where language dominance and exposure were quantified via a background questionnaire. During the task, the children recognized both English and Spanish words embedded in sentences in their non-dominant language but when the words were embedded in sentences in the dominant language, they only recognized the dominant language words. Finally, infants aged 0;10 who were bilingual in French and a second language employing lexical stress could discriminate stress contrasts more quickly if they were *less* dominant in French—a language that does not employ lexical stress (Bijeljac-Babic et al. 2012). Taken together, these studies demonstrate that the quantity of bilingual children's exposure to their two languages predicts speech-language outcomes, including speech processing, perception, and vocabulary size.

2.4 Estimating dual language exposure from daylong recordings

As reviewed above, dual language exposure is typically quantified from parental report in diaries or via interviews and questionnaires. But this exposure can also be estimated using naturalistic language samples from daylong audio recordings. One example is Orena et al. (2019) who collected three different daylong recordings, on separate days, from French-English infants (aged 0;10). Every other 30-second clip from the recordings was annotated for language (French, English, or Mixed) and speaker (target child, father, mother, or sibling). In-person bilingual language use interviews were also conducted with caregivers. Dual language estimates from the daylong recordings and parental interviews were compared. Estimates from the two techniques were broadly similar, though comparisons across the three days of recording showed considerable variation in language exposure within participants. The authors suggest that both naturalistic and parental self-report estimates are important for accurate dual language exposure estimation.

Elsewhere, a combination of manual and automated annotation of daylong recordings was employed to correlate dual language exposure with vocabulary outcomes (Ramírez-Esparza et al. 2017). In that study, forty 30-second clips were drawn from daylong recordings made of Spanish-English infants at 0;11. To ensure the presence of adult speech, only clips containing high algorithmically-derived adult word counts (using LENA software) were selected and manually annotated for language and speech register (various types of CDS were coded; see article for details). The naturalistic estimates of CDS in each language correlated positively with vocabulary scores at 2;0.

Finally, Marchman et al. (2017) also used manual and automated analyses to code 5-minute samples from daylong recordings of young (aged 3;0) Spanish-English bilinguals for 1) speech register (CDS or adult-directed speech) and 2) language, determined by the proportion of Spanish and English words in the clip. Then, the quantity of LENA-estimated adult words in each CDS clip was multiplied by the proportion of Spanish or English in each clip, as judged by the human coders, to determine CDS quantities in Spanish and English. Children's dual language exposure was also measured via parental report. CDS estimates from the daylong recordings predicted children's performance better than parental report on measures of Spanish-English vocabulary and looking-while-listening speech processing tasks, suggesting that naturalistic estimates of dual language exposure could be more valid than parental report.

It should be noted that the bilingual populations studied in Marchman et al. (2017) and Orena et al. (2020) differ in more than just language pairings: English and French are both high-prestige languages in Montreal, whereas Spanish-English bilingualism does not have the same level of acceptance in the United States. Given the different findings between these two studies, it is conceivable that caregiver self-reports may be less reliable when caregivers face discrimination for their language practices.

Furthermore, researchers and practitioners have rarely considered the difficulty of extending parental report methods—including questionnaires—to populations who do not frequently participate in behavioral research. Caregivers with low levels of literacy, for example, might have difficulty completing language diaries or other written self-reports that have to be completed over the course of a child's day (e.g. Carbajal & Peperkamp 2019; Place & Hoff 2011). Similarly, in-person oral interviews (e.g. Marchman et al. 2017) may be more difficult for populations that are unfamiliar with behavioral research methods, or face stigmatization for speaking one of their languages. The difficulty in carrying out traditional bilingual research methods in under-studied populations may actually inhibit the inclusion of these groups in behavioral research. Yet including under-studied bilingual populations is essential to understanding how bilingual and monolingual language learning experiences differ and how bilingual development differs from one sociolinguistic context to the next.

3 Current Study

Recognizing the need for ecologically-valid estimates of children's language exposure, but also the practical difficulties of annotating the naturalistic data collected in these studies, this study proposes a dataflow that samples from daylong recordings to efficiently estimate the proportion of CDS and dual language exposure in a child's day. Previous research that sampled from daylong audio recordings to estimate children's language exposure took conscious samples, for example high-volubility clips from morning, afternoon, and evening in the recording (Orena et al. 2019, Ramírez-Esparza et al. 2017). Here, we instead use clips randomly sampled from each recording. An advantage of this approach is that it does not rely on estimates of volubility and can therefore be used with longform recordings collected via a variety of methods, not only LENA.

A key question for assessing the utility of a sampling approach is how much annotation is necessary to obtain reliable estimates of the desired speech and language categories. The amount of annotation necessary will depend on the overall frequency of the category we want to estimate (e.g., how much Spanish a child hears in a day) (Micheletti et al. 2020). Because this is not known in advance, the current study proposes a data-driven approach. Specifically, we use an application that enables annotators to monitor how estimates of language/speech register change over the course of annotation. We use this to empirically determine the number of randomly sampled clips from which to base our estimates on a child-by-child basis. Having done that, we then take multiple steps to validate the estimates derived from this random sampling approach.

The daylong recordings analyzed here come from bilingual communities in the United States (Spanish-English) and Bolivia (Quechua-Spanish). Examining two different bilingual communities with different histories and relations to the dominant language and culture (one immigrant community and one indigenous community) lends confidence that this workflow may be used across various cultural, linguistic, and sociopolitical contexts. However, we also acknowledge that the workflow may look different across socio-cultural settings, and we encourage others to use this application on samples from diverse communities to evaluate the number of annotated clips necessary to get reliable CDS and dual language estimates, in addition to other characteristics of children's language environments.

We make the following hypotheses:

1. Employing random sampling, we will efficiently estimate the proportion of speech in different speech registers and languages in children's environments. Stable estimates between speech-language categories will be met after limited manual annotation.

Having established that stable estimates of the CDS and language proportions can be achieved by random sampling, we will validate this approach by comparing random sampling estimates to more time-intensive, gold standard estimates made by annotating every other clip from the recording, or ALL-DAY SAMPLING, as in Orena et al. (2019).

2. The random and all-day sampling methods will produce similar estimates of the proportion of each speech register and language category, suggesting that annotations of relatively small amounts of data based on random sampling is an efficient way of estimating ambient language characteristics from daylong recordings.

Research questions 1 and 2 are tested in both bilingual communities, Spanish-English in the United States and Quechua-Spanish in Bolivia, for all speech registers and languages present in the children's environments. The study predictions are the same across both sites and both speech-language categories.

A final step to validate this approach is to compare random and all-day sampling estimates of dual language exposure based on daylong recordings to estimates derived from parental report. Extensive parental reports of dual language exposure were collected via oral interviews from Spanish-English families in the United States, but not for the Quechua-Spanish communities in Bolivia. Thus, only bilingual Spanish-English language exposure in the United States is evaluated in this way.

3. It is possible that the bilingual language exposure estimates made via random/all-day sampling will be similar to estimates derived from parental report. This result would suggest that these methods provide compatible estimates of children's dual language exposure. Alternatively, if the random sampling and parental questionnaire estimates of dual language exposure differ, this may suggest that 1) naturalistic samples estimate some aspects of children's daily dual language exposure better than parental report in some communities, and/or 2) a single daylong recording does not adequately reflect children's entire dual language exposure.

4 Method

4.1 Participants

Daylong audio recordings of 10 infants from two bilingual child language corpora were used: N=5 from an immigrant Spanish-English corpus collected in the United States (Weisleder & Mendelsohn 2019) and N=5 from a Quechua-Spanish corpus collected in and around a mid-size town in southern Bolivia (Cychosz 2018). Infants were age-matched across corpora (Spanish-English: *M*=8.94, *SD*=2.54; Quechua-Spanish: *M*=8.6 mos, *SD*=2.61). See Table 1 for complete demographic information. Given differences in educational opportunities for women across the two sociocultural contexts, participants were not matched for maternal education. However, in both groups, most mothers had less than a high school education. All caregivers reported normal speech and hearing development for their infants. Data collection and all secondary analyses conducted here were approved by Institutional Review Boards at New York University School of Medicine, Northwestern University, and the University of California, Berkeley.

Table 1
Participants' demographic information

Corpus	**Infant age (mos)**	**Gender**	**Maternal education (yrs)**	**N of older siblings**	**Maternal language dominance**
Quechua-Spanish	5.7	F	3	2	Bilingual Quechua-Spanish
	6.0	M	12	2	Bilingual Quechua-Spanish
	9.5	F	5	3	Quechua-dominant
	9.8	M	5	3	Bilingual Quechua-Spanish
	12.3	F	6	2	Bilingual Quechua-Spanish
Spanish-English	6.5	M	7	1	Spanish-dominant
	7.2	F	9	3	Spanish-dominant
	7.9	M	12	1	Spanish-dominant
	10.5	F	11	1	Bilingual Spanish-English
	12.6	M	13.5	1	Spanish-dominant

4.2 Data collection

To collect the audio recordings, each infant wore a small, lightweight (< 60 g, 5.5 x 8.5 x 1.5 cm) LENA digital language processor (i.e., audio recorder) in the pocket of a specialized shirt or vest (Greenwood et al. 2011). Parents were instructed to record over the course of a typical day, aiming for 12-16 hours. All families in the Spanish-English corpus completed 16-hour recordings. The average recording length in the Quechua-Spanish corpus was 10.47 hours (*SD*=3.23; range=7.77-16). The recordings in the Spanish-English corpus typically captured an entire day from morning to night, while the recordings in the Quechua-Spanish corpus sometimes started later in the day. The shorter recording lengths in the Quechua-Spanish corpus were due to the distance that the researcher had to travel each day to deliver recorders to the families and because families were instructed to pause the recording throughout the day if they wanted.

Each participating family also completed a bilingual language use questionnaire. For the Quechua-Spanish-speaking families in Bolivia, a researcher carried out a brief oral survey with the primary caregiver that included questions about the central caregivers' and both sets of grandparents' language use, and more specific information about the primary caregiver's code-switching habits. The Spanish-English-speaking families in the United States completed an oral survey based on the Bilingual Background Interview (Marchman & Martinez-Sussmann 2002). Mothers were first asked about people with whom the child had regular weekly contact. This was followed by questions asking how many hours per week the child spent with each person and what language(s) the person spoke with the child (English, Spanish, or both). When both languages were reported, mothers were asked to estimate the relative proportion of English and Spanish used.

4.3 Data workflow

A series of custom Python scripts were written to process the recordings (available at `https://github.com/megseekosh/Categorize_app_v2`). This entire dataflow process is illustrated in Figure 1. First, each recording was segmented into 30-second clips. The 30-second clip length has been used in previous work that has sampled and annotated portions of children's daylong recordings to estimate dual language (Orena et al. 2019; Ramírez-Esparza et al. 2017) and CDS exposure (Ramírez-Esparza et al. 2014), among other behavioral variables (Micheletti et al. 2020). There were, on average, 1257 clips per recording in the Quechua-Spanish corpus (*SD*=387; range=933-1920). There were 1920 clips per recording for all recordings in the Spanish-English corpus.

Once the recordings were segmented, a standard vocal activity detector (Usoltsev 2015) was run over all of the clips and the percentage of the clip containing vocal activity was reported. In practice, the vocal activity metric served two purposes. First, clips that contained 0% vocal activity were not drawn for annotation. Second, prior to annotation, the first and second authors listened to portions of the recordings containing extended stretches of low vocal activity to determine if the infant was sleeping. The researchers determined if the infant was sleeping by listening for relative quiet in the background, lack of vocalizations from the target infant, and heavy breathing or snoring. If the researcher found that a clip contained audio of a sleeping infant, the clip was marked not to be drawn for annotation.

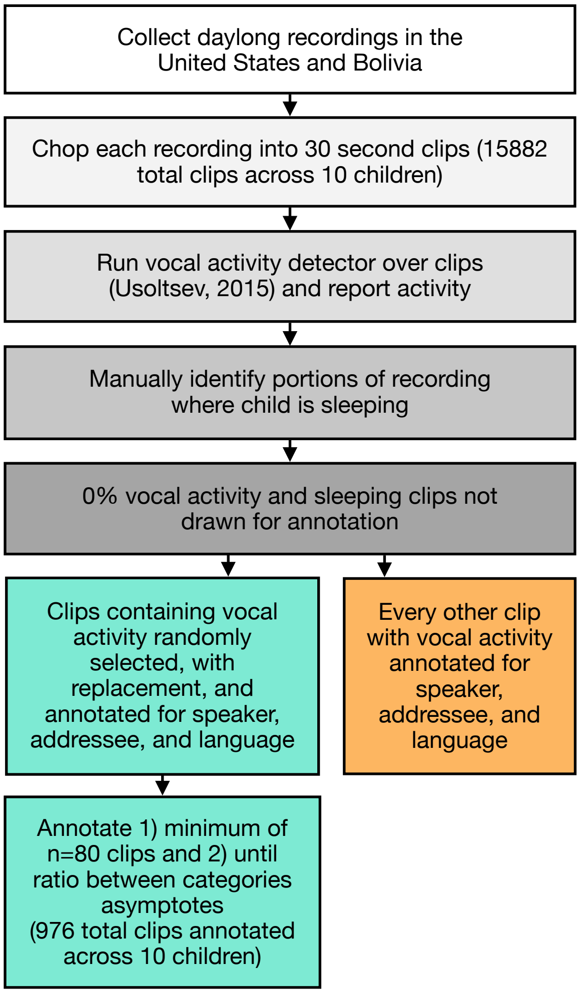

Collect daylong recordings in the United States and Bolivia
Chop each recording into 30 second clips (15882 total clips across 10 children)
Run vocal activity detector over clips (Usoltsev, 2015) and report activity
Manually identify portions of recording where child is sleeping
0% vocal activity and sleeping clips not drawn for annotation
Clips containing vocal activity randomly selected, with replacement, and annotated for speaker, addressee, and language
Every other clip with vocal activity annotated for speaker, addressee, and language
Annotate 1) minimum of n=80 clips and 2) until ratio between categories asymptotes (976 total clips annotated across 10 children)

Figure 1. Audio clip generation, selection, and annotation workflow. Green boxes indicate workflow for random sampling method, orange box indicates workflow for all-day sampling method, gray boxes indicate workflow for both methods.

4.4 Data annotation

The first and second authors annotated the Spanish-Quechua and Spanish-English recordings, respectively. The first author is fluent in Spanish and familiar with Quechua, has done fieldwork in Bolivia, and knows the families who completed the recordings, all of which facilitated annotation. The second author is a native Spanish-English bilingual and had access to logs from the day of recording indicating the number of adults and children that interacted with the child throughout the day. To facilitate annotation of the 30-second clips, a custom Generalized User Interface (GUI) application was built. (See `https://github.com/megseekosh/Categorize_app_v2` for further details on GUI application structure and instructions for use.) The GUI application led the annotator through the steps of clip annotation. First, a 30-second clip was selected from a participant's recording. To speed up annotation, clips were drawn but not annotated if they had 0% vocal activity or if the child was determined to be sleeping. Clips that contained speech from a researcher (for example when the recording was turned on as the researcher was leaving the participant's home) were also identified prior to annotation and were likewise drawn but not annotated.

After drawing a clip, the annotator would listen to the clip and first ensure that there was speech within the clip. If so, the annotator chose between the following options for speaker, addressee (speech register), and language:

1. **Speaker**: Adult (Adult Female, Adult Male, Multiple Adults), Other Child, or Unsure

2. **Addressee**: Target Child, Other Child, Adult or Unsure

3. **Language**: Quechua/English, Spanish, Mixed, or Unsure

Annotators coded the language of conversation between all combinations of speaker and addressee present in the clip (e.g. Other Child speaking Spanish to Target Child, Adult speaking in Quechua to another adult, Adult using Mixed language with another child). For speaker and addressee annotation, the 'Target Child' was the infant wearing the recorder and 'Other Child' was defined as any individual who had not gone through puberty. Annotators could usually determine whether a speaker was a child or an adult because they had information on the household members and their ages. However, when an annotator could not recognize a speaker's voice, the speaker was labelled as 'Other Child' if their voice sounded pre-pubescent.

For language annotation, if only Spanish, Quechua, or English was spoken for a particular combination of speaker and addressee (regardless of the quantity of speech), the researcher marked the appropriate language category. If a speaker in the clip used both Quechua and Spanish to a particular addressee type—either code-switching within a sentence or two separate conversations—the interaction was marked as 'Mixed.' However, for example, if a speaker in a clip used Spanish to speak to a child but English to speak to an adult, the annotator would choose ‘Spanish’ for child-directed speech and ‘English’ for adult-directed speech. Note that this means that a single clip could have multiple annotations.

For those clips where the speaker, addressee, or language was not clear (e.g., when a speaker uses a name, expression, or a non-language vocalization such as cooing sounds, hushing sounds, nonsense syllables, whistling, humming, or laughing), annotators could select ‘Unsure.’ Since speaker, addressee, and language were coded separately, cases in which the annotator could not determine one category were still coded for the remaining categories; for example, if the addressee was marked as ‘Unsure’, speaker and language were still coded. In practice, the ‘Unsure’ annotation was used most often for clips where a conversation was taking place in the background, far from the child, making it difficult to determine the language, speaker, or addressee categories. Finally, if a particular category of speaker and addressee did not occur (e.g., there was no adult speech to children), this category was left blank.

The proportions of language and speech register categories in each recording were determined by dividing the number of speech register/language annotations containing each category by the total number of annotations. We chose total number of *annotations* as our denominator, instead of total number of clips, because some clips had multiple conversations between different speakers using different languages (as described above). Thus, for example, if a clip contained CDS in Spanish and adult-directed speech in English, that clip would contribute 1 Spanish annotation and 2 total annotations to the estimate of proportion Spanish. The clip would also contribute 1 CDS annotation and 2 total annotations to the estimate of proportion CDS.

Media presence, including the language of the media (Spanish, Quechua/English, Mixed, Unsure, no language) was also annotated. Language exposure estimates in this paper do not factor in language exposure from media. Finally, annotators made the following binary decisions about the clip:

1. Is the target child vocalizing?

2. Is the target child sleeping?

3. Is there personal identifying information in the clip?

See `https://github.com/megseekosh/Categorize_app_v2/blob/master/FAQs.MD` for further details on annotation decisions, including a list of frequently asked questions used to standardize annotation between research personnel.

4.5 Methods of annotation

The 30-second clips from the daylong recordings were drawn in two ways: by sampling randomly from the recording (random sampling) and by selecting every other clip (all-day sampling). To sample randomly from the recording, a script randomly drew 30-second clips for annotation, with replacement. Clips with 0% vocal activity or marked as 'Child Sleeping'/'Researcher Present' were drawn but not annotated. For the all-day sampling annotation method, clips were drawn chronologically from the beginning of the recording to the end, skipping every other clip. Again, clips with 0% vocal activity or marked as 'Child Sleeping'/'Researcher Present' were drawn but not annotated.

Annotation using the random sampling method was always conducted first. After random sampling annotation had been completed for all participants, and following at least a 2-month period, the same coders annotated the same recordings using the all-day sampling method. This order of annotation is important because it ensures that the estimates derived from random sampling resemble those that would be obtained in real life annotation scenarios, where annotators would not have the benefit of having listened to the whole recording when annotating via random sampling.

For random sampling, as annotators drew and listened to the 30-second clips, they simultaneously ran a Jupyter notebook (included in the Github repository) that quantified progress towards annotation. Specifically, after each clip was annotated, the proportion of language (Quechua/English, Spanish, or Mixed) and speech register (Target child-directed speech, Adult-directed speech, Other child-directed speech) was updated for that recording. For each recording, annotation was cut off when both a quantitative *and* a qualitative criterion were met. The quantitative criterion was that a minimum of n=80 clips from the recording had to be annotated for language and/or speech register; the qualitative criterion was that the estimated proportion of each language and speech register categories had asymptoted (as exemplified in Figures 2 and 3), suggesting that a stable estimate had been reached. The n=80 criterion was used to ensure that estimates were not made on the basis of too few clips, even if stability appeared to have been reached earlier. The qualitative asymptoting criterion was used because the number of samples needed to obtain a reliable estimate varies as a function of the category's frequency distribution (Micheletti et al. 2020), which is not known in advance. By monitoring the cumulative estimates as annotation progressed, annotators could make data-driven decisions about when to cut off annotation.

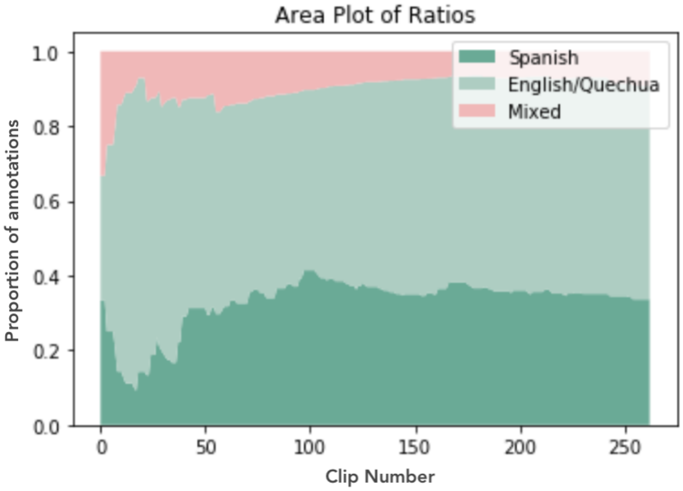


Figure 2. Example area plot of language proportion by number of clips annotated (Participant #1032 from the Quechua-Spanish corpus). Area plots were used to track progress towards language estimate stability during daylong recording annotation.

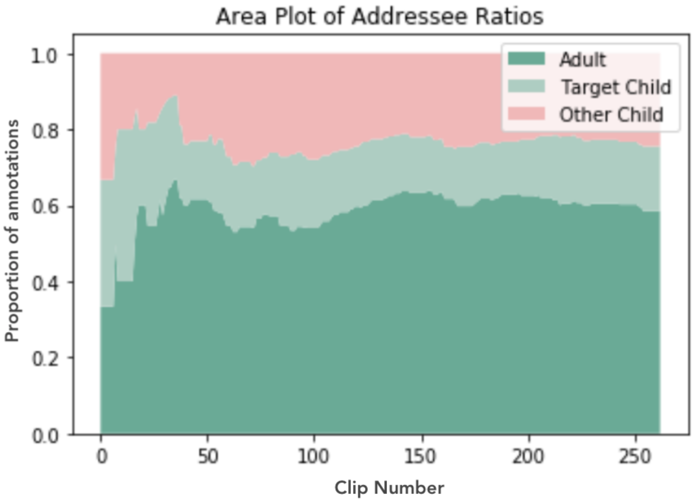


Figure 3. Example area plot of addressee proportion by number of clips annotated (Participant #1032).

Table 2 displays the percentage of overall clips from each recording that were drawn during random and all-day sampling, as well as the percentage of those drawn clips that were annotated for language and/or speech register (those that contained speech from adults or other children). See Appendix A for a table with the percentages of each recording annotated. Fewer clips were annotated than drawn because the drawn clips include those without any speech, with only electronic noise, etc. The percentage of clips requiring annotation for language and speech register in order to reach category stability for the random sampling method ranged from 4-11% of total clips from the recording and 8-14% of clips containing speech. Therefore, if the estimates derived from random sampling are indeed similar to those derived from all-day sampling, random sampling has the potential to save large amounts of annotation time.

5 Results

Our primary research question asks if random sampling of daylong recordings can accurately estimate infants' CDS and dual language exposure. In the following section, we evaluate this question by first demonstrating how language and speech register categories stabilized during annotation of randomly selected segments. Then, we correlate estimates derived from random sampling with those made from all-day sampling and with parent report. We complement these correlations with a series of simulations estimating the proportion of language and speech registers in each recording from different random samples.

All analyses were conducted in the RStudio computing environment (version: 1.3.1073; RStudio Team 2020). Data visualizations were created with ggplot2 (Wickham 2016). Modeling was conducted using a combination of the lme4 and lmerTest packages (Bates et al. 2015; Kuznetsova et al. 2017). All scripts to replicate these analyses are publicly available in the project's GitHub repository.

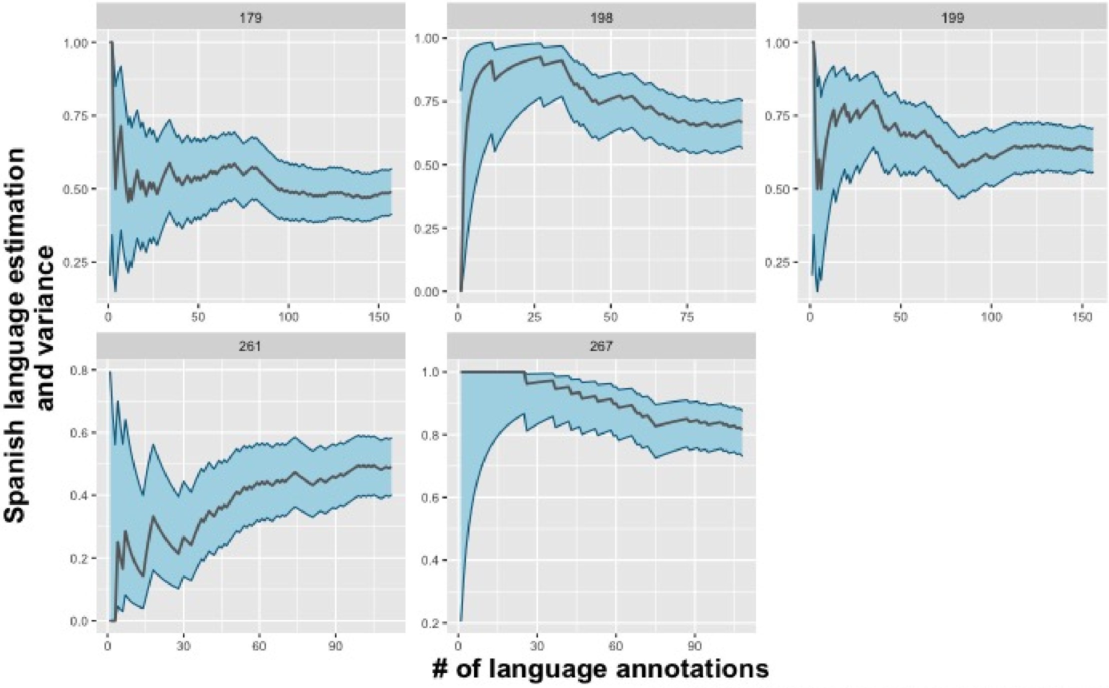
Table 2
Number of clips drawn and number of clips annotated for language/speech register, by child and annotation method.

	# of clips drawn (% of total clips)		# of clips annotated (% of available clips)	
Corpus (ID)	Random	All-day	Random	All-day
Spanish-English (267)	345 (17.97 %)	960 (50 %)	101 (13.10 %)	274 (35.54 %)
Spanish-English (261)	290 (15.10 %)	960 (50 %)	92 (8.18 %)	294 (26.13 %)
Spanish-English (199)	192 (10.00 %)	960 (50 %)	118 (10.61 %)	467 (42.00 %)
Spanish-English (198)	284 (14.79 %)	960 (50 %)	81 (7.96 %)	302 (29.67 %)
Spanish-English (179)	192 (10 %)	960 (50 %)	120 (8.05 %)	633 (42.48 %)
Quechua-Spanish (1081)	249 (20.31 %)	613 (50 %)	92 (13.83 %)	285 (42.86 %)
Quechua-Spanish (1077)	137 (11.93 %)	574 (50 %)	83 (8.15 %)	355 (34.84 %)
Quechua-Spanish (1075)	267 (28.65 %)	466 (50 %)	81 (14.21 %)	199 (34.91 %)
Quechua-Spanish (1060)	154 (14.58 %)	528 (50 %)	111 (12.01 %)	405 (43.83 %)
Quechua-Spanish (1032)	263 (13.70 %)	960 (50 %)	97 (10.16 %)	372 (38.95 %)

'Available' clips refer to those where the child was not sleeping, the researcher was not present, and speech was detected by the vocal activity detector. The 'Percentage of available clips' is the number of clips that were annotated divided by the total number of available clips. 'Number of clips annotated' does not include those clips only containing target child vocalizations and/or media.

5.1 Stabilization of language and speech categories from random sampling

Our first research question asks if random sampling results in stable estimates of the proportion of CDS and Spanish/Quechua after a modest amount of annotation. (For the remainder of the results, CDS refers to target child-directed speech, and excludes other child-directed speech such as that directed to the target child's siblings.) First, we qualitatively assessed that a stable estimate was reached by observing 1) stabilization in the mean of the estimate and 2) a reduction in the variance of the estimate as a function of the number of annotations made. Since both the language and speech register outcomes were binary variables (e.g. +/- Spanish, +/- CDS), we use the Wilson binomial confidence interval as an indicator of variance around the estimate. Figures 4-6 illustrate how, as more annotations are made, 1) the category estimate stabilizes (gray line) and 2) the variance surrounding the category estimate decreases and stabilizes (blue ribbon).



'# of language annotations' includes 'unsure' language annotations.

Figure 4. Variance in Spanish language estimation as a function of annotations made: U.S. corpus. The gray line represents the estimate of the proportion of Spanish in the recording and the blue ribbon represents a 95% Wilson binomial proportion confidence interval. Note that these visualized estimates are calculated over all clips containing language, and include ‘unsure’ annotations, so they may differ slightly from estimates reported in Section 5.2.

We next evaluated this approach by fitting a series of linear mixed effects models—for language and for speech register—to test the extent to which clip number predicted a reduction in the variance estimates from each recording. The outcome for the language models was variance (confidence interval range) of the Spanish language estimation and the outcome for the speech register models was variance of the CDS estimation. The language models were fit to Spanish language estimates from both corpora, instead of Quechua estimates from the Quechua-Spanish corpus and Spanish estimates from the Spanish-English corpus, to increase sample size. For each outcome, four models

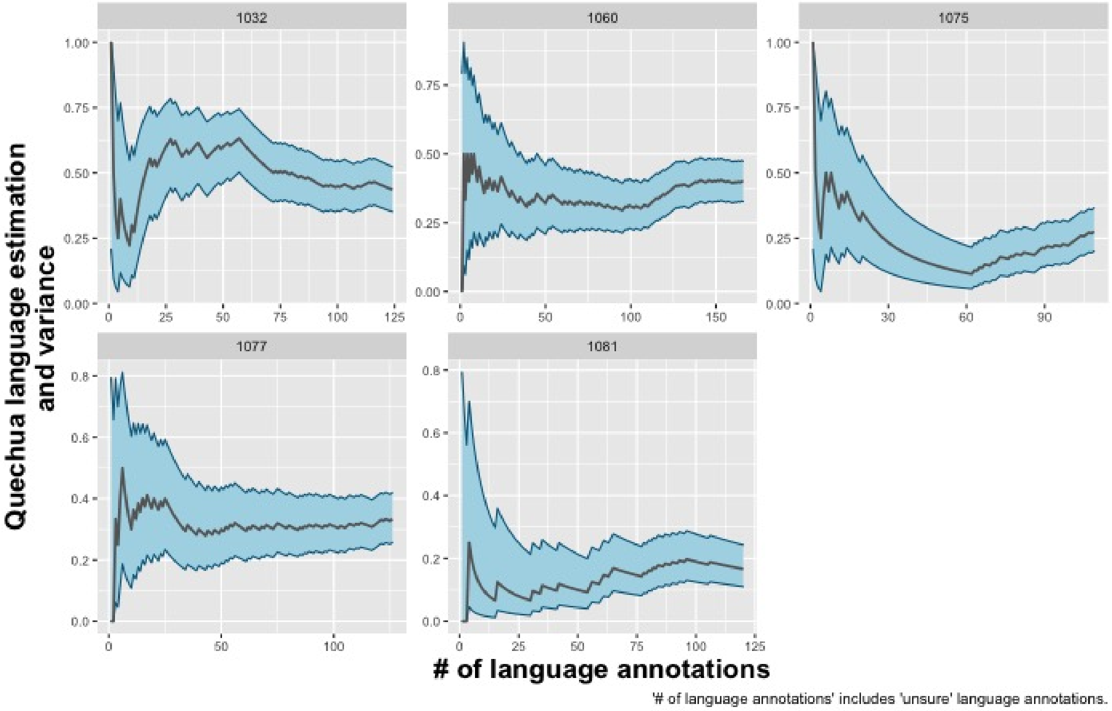
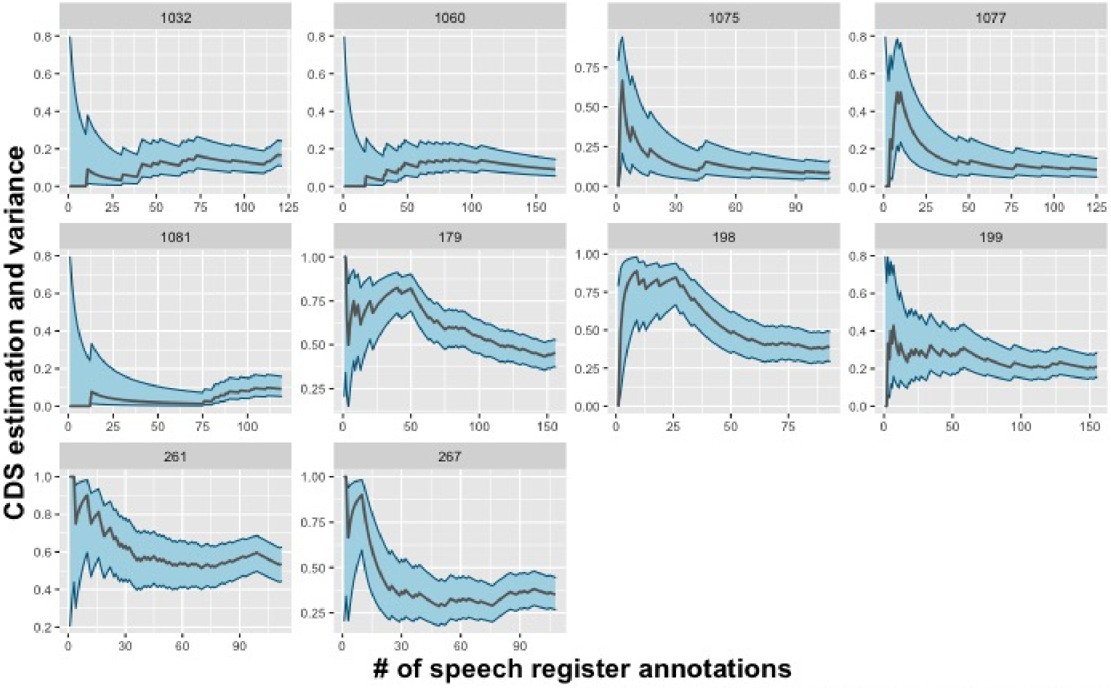


Figure 5. Variance in Quechua language estimation as a function of annotations made: Bolivia corpus. The gray line represents the estimate of the proportion of Quechua in the recording and the blue ribbon represents a 95% Wilson binomial proportion confidence interval. Note that these visualized estimates are calculated over all clips containing language, and include 'unsure' annotations, so they may differ slightly from estimates reported in Section 5.2.

were fit: over the first, second, and third of random annotations made in each recording, and over the all-day annotations.

The logic for this analysis is that the effect of annotation number in predicting change in the estimate's variance will depend upon the number of annotations that have been factored in. The slope should be steepest for the first third of randomly-made annotations, where increased annotation has an outsize effect upon the outcome variable, but should be increasingly shallower over the second and third third of randomly-made annotations. Finally, the slope fit over the all-day annotations should be the shallowest, indicating that the additional annotations in all-day sampling (relative to the subset used in random sampling) have a smaller effect in reducing the variance around the estimate. The differing slopes across these datasets would suggest that there are diminishing returns to continued annotation after a certain point.

Each model included a random effect of participant. The predictor **Annotation Number**



'# of speech register annotations' includes 'unsure' speech register annotations.

Figure 6. Variance in target child-directed speech estimation as a function of annotations made. The gray line represents the estimate of the proportion of CDS in the recording and the blue ribbon represents a 95% Wilson binomial proportion confidence interval. Note that these visualized estimates are calculated over all clips containing language, and include 'unsure' annotations, so they may differ slightly from estimates reported in Section 5.2.

significantly improved all model fits, always with a negative beta coefficient, suggesting that increased annotation reduced the amount of variability in language/speech register estimation over all datasets. However, as predicted, the slope in variance reduction of Spanish language estimation was steepest over the first third of annotations made (β=-.00971, t=-33.93, p<.001), followed by the second third (β=-.00126, t=-32.59, p<.001), third third (β=-.00075, t=-68.96, p<.001), and finally the all-day annotations (β=-.00034, t=-59.49, p<.001). The same pattern was present for the slopes for variance in CDS estimation where the steepest slope was over the first third of annotations (β=-.00958, t=-31.57, p<.001), followed by the second third (β=-.00137, t=-41.45, p<.001), and third third (β=-.00067, t=-33.04, p<.001). Finally, the slope was shallowest over the all-day annotations (β=-.00029, t=-51.92, p<.001). Together, these modeling results suggest that increased predictor units (annotations) are having increasingly smaller effects upon the outcome variable (running variance in category estimation).

Overall, this analysis first demonstrated that running variance decreases, and the category proportion estimate stabilizes, over the course of annotation for both language and speech register in both corpora. The modeling supported these observations, demonstrating that each additional annotation predicts a reduction in the variance around the estimate and that this reduction is progressively shallower as annotation continues. Although these results do not tell us the *optimal* point to cease annotation, they suggest that relatively stable estimates can be reached after a modest amount of annotation. By monitoring this process of stabilization, the current study derived estimates from random sampling based on an average of 49 minutes per recording, or 11% percent of clips containing speech (see Table 2). Next, we evaluate how well these estimates approximate estimates obtained from all-day sampling.

5.2 Comparing random and all-day annotation

Next we compared estimates for the proportion of each language and speech register category derived from random and all-day sampling for both corpora. Figures 7 and 8 show the proportion of each language and speech register per participant as estimated by each method.

Focusing first on the language categories, Figure 7 shows there is substantial within-group variability in the proportion of the minority language heard by children (Quechua for the Bolivia sample, Spanish for the U.S. sample). To quantify differences in the estimates derived from each annotation method, we computed the absolute error between minority language proportions made via random sampling (estimated value) and all-day sampling (true value). Although the average absolute errors between random and all-day minority language estimates were slightly higher in the Quechua-Spanish corpus (M=0.05) than the Spanish-English corpus (M=0.04), these relatively low errors indicates that the estimates derived from random sampling closely approximated those obtained from the more time-intensive method.

Next, we computed correlations between estimates of the minority language derived from random and all-day sampling. These correlations were strong for Spanish in the Spanish-English corpus (r(5)=.96, p=.01; Figure 9) and for Quechua in the Quechua-Spanish corpus (r(5)=.90, p=.04; Figure 10). The strong correlations indicate that random sampling effectively captured variability in infants' exposure to the minority language as captured by the daylong recordings.

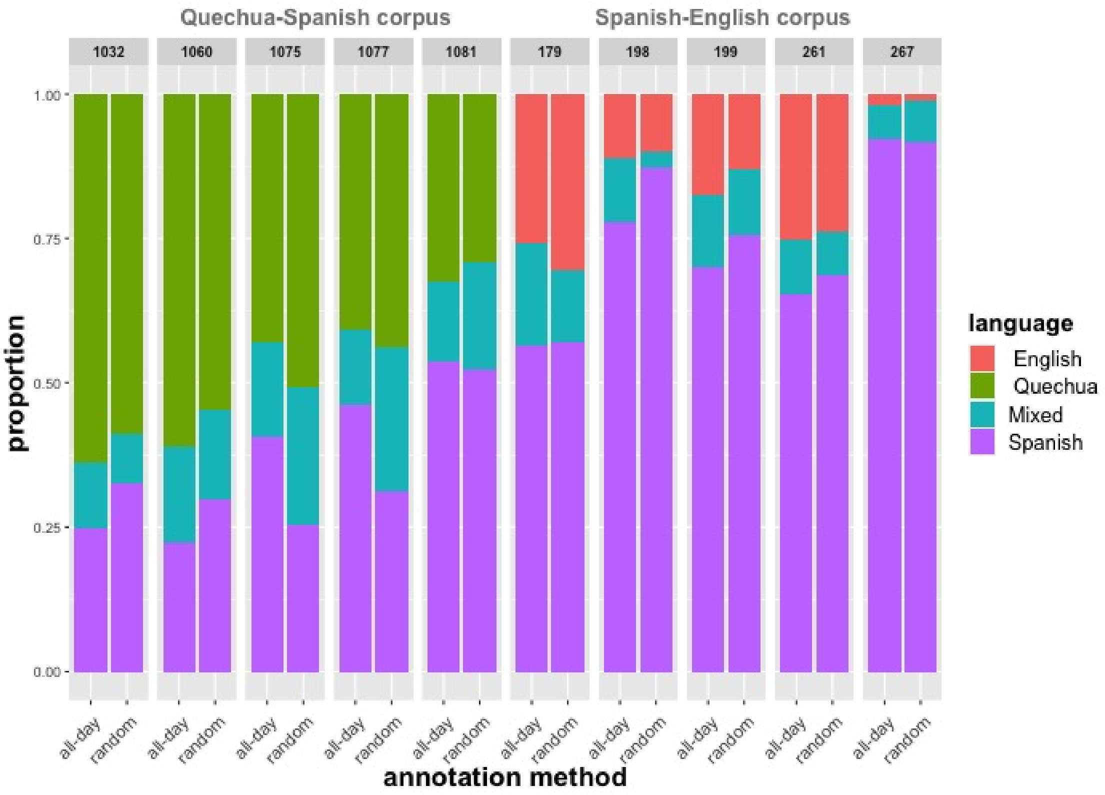


Figure 7. Proportion of language categories, by child and annotation method.

Table 3
Average minority language estimates by corpus and annotation method.

Corpus (ID)	Annotation Method		Absolute error	n=100 simulations of random sampling Avg. (SD) Range	Within-corpus correlation between random and all-day estimates	Avg. & Range of within-corpus correlation between simulated and all-day estimates
	Random	All-day				
Spanish-English (179)	0.57	0.57	0.00	0.56 (0.04) 0.49 - 0.68	r=0.96, p=0.01	0.97, 0.87-1
Spanish-English (198)	0.87	0.78	0.09	0.78 (0.04) 0.69 - 0.87		
Spanish-English (199)	0.76	0.70	0.06	0.70 (0.04) 0.61 - 0.81		
Spanish-English (261)	0.69	0.65	0.04	0.65 (0.04) 0.56 - 0.75		
Spanish-English (267)	0.92	0.92	0.00	0.93 (0.02) 0.88 - 0.99		
Quechua-Spanish (1032)	0.59	0.64	0.05	0.63 (0.04) 0.51 - 0.74	r=0.90, p=0.04	0.96, 0.68-1
Quechua-Spanish (1060)	0.55	0.61	0.06	0.60 (0.04) 0.51 - 0.69		
Quechua-Spanish (1075)	0.51	0.43	0.08	0.44 (0.06) 0.27 - 0.58		
Quechua-Spanish (1077)	0.44	0.41	0.03	0.40 (0.04) 0.29 - 0.50		
Quechua-Spanish (1081)	0.29	0.32	0.03	0.33 (0.04) 0.25 - 0.45		

Focusing next on the speech register categories, Figure 8 shows differences in the proportion

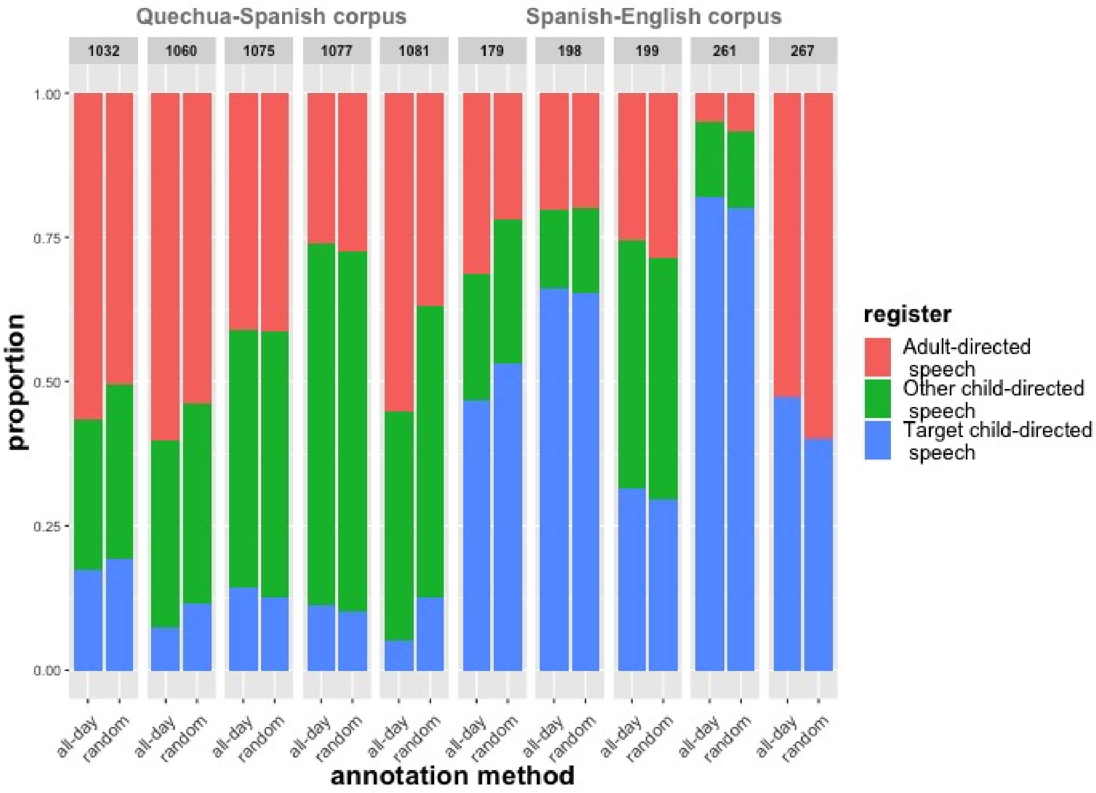


Figure 8. Proportion of speech register categories, by child and annotation method.

of speech directed to the target child (CDS) as derived from random and all-day sampling for participants across the two corpora. We again compared estimates of CDS for the two annotation techniques by computing the absolute error (Table 4). This metric shows that the estimates of CDS are similar across annotation methods.

Next, we computed correlations between estimates of CDS from random and all-day sampling. The correlation computed over the two corpora combined was very strong (r(8)=.99, p<.001). The correlation between random and all-day sampling estimates for just the Spanish-English corpus was again strong (r(5)=.97, p=.01; Figure 11). The correlation for just the Quechua-Spanish corpus was notably weaker (r(5)=.64; p=.24).

One possible explanation for the weaker correlation in the Quechua-Spanish corpus is that the range of CDS exposure estimates in that corpus was quite small (0.10-0.19 proportion CDS from random sampling and 0.05-0.17 proportion CDS from all-day sampling), so the correlation may be limited by range. Indeed, the range of CDS estimates was much larger in the Spanish-English corpus

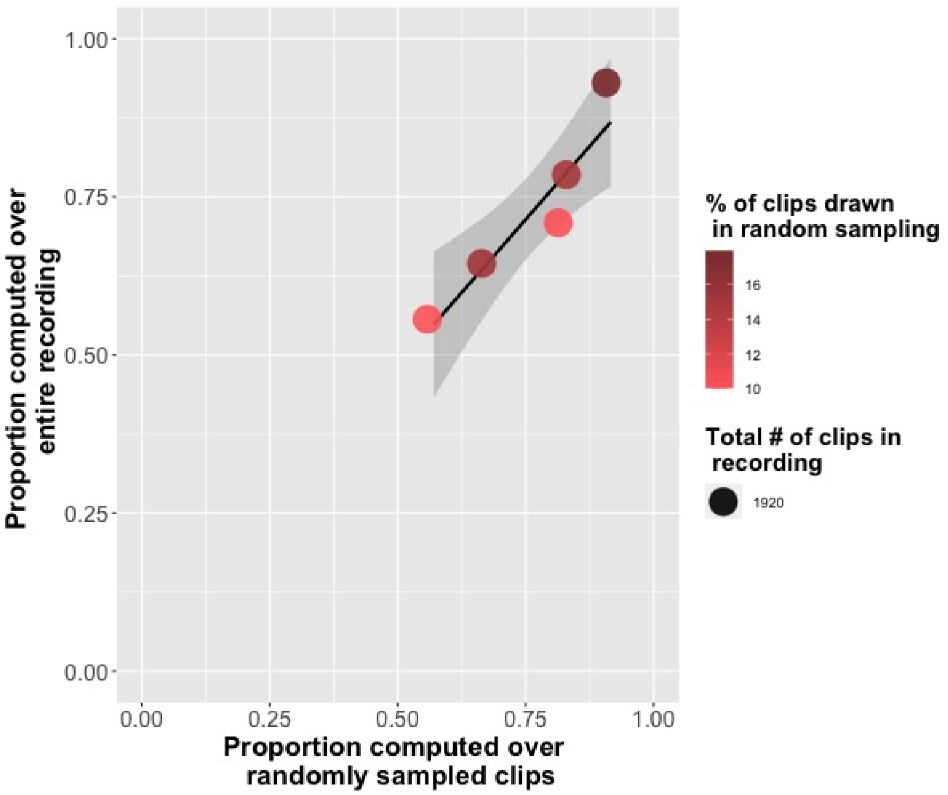


Figure 9. Proportion of Spanish clips in U.S. corpus, by annotation method.

for both random (0.29-0.80 proportion CDS) and all-day sampling (0.31-0.82 proportion CDS), as Figure 11 clearly shows. In addition, the proportion of CDS in the Quechua-Spanish recordings was overall much lower than the proportion of CDS in the Spanish-English recordings, leading us to hypothesize that it may be more difficult to reliably estimate low-frequency events using random sampling (see also Micheletti et al. 2020).

To address the possibility that low-frequency events may lead to erroneous estimates, we calculated the correlation and absolute error between random and all-day sampling estimates of the more frequent speech register category in the Quechua-Spanish corpus: adult-directed speech. The correlation between annotation methods for adult-directed speech in the Quechua-Spanish corpus was stronger than for CDS, approaching significance (r(5)=.84; p=.08), with an average absolute error of 0.07 (*SD*=0.07). This result suggests that estimates of high-frequency events made via random sampling may be more reliable than estimates of low-frequency events. We return to this in the Discussion.

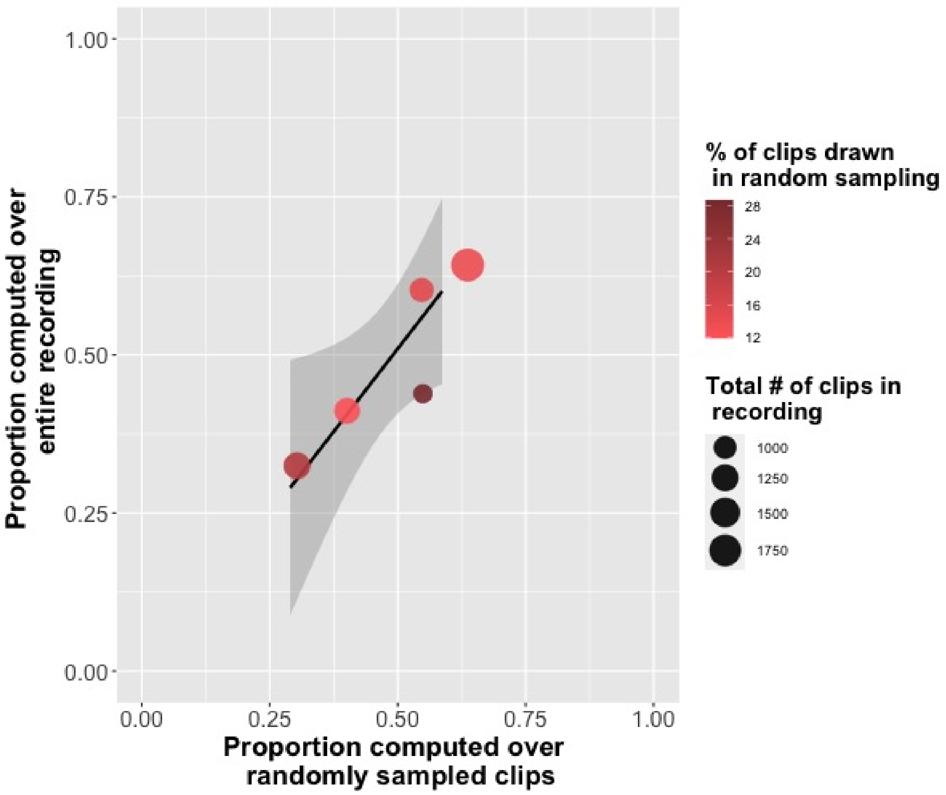


Figure 10. Proportion of Quechua clips in Bolivian corpus, by annotation method.

Table 4

Target child-directed speech estimates by child and annotation method.

	Annotation Method					
Corpus (ID)	Random	All-day	Absolute error	n=100 simulations of random sampling Avg.(SD) Range	Within-corpus correlation between random and all-day estimates	Avg. & Range of within-corpus correlation between simulated and all-day estimates
Spanish-English (179)	0.53	0.47	0.06	0.47 (0.04) 0.38-0.55	r=0.97 , p=0.01	0.98, 0.89-1
Spanish-English (198)	0.65	0.66	0.01	0.66 (0.06) 0.47-0.80		
Spanish-English (199)	0.29	0.31	0.02	0.31 (0.04) 0.22-0.42		
Spanish-English (261)	0.80	0.82	0.02	0.82 (0.04) 0.73-0.88		
Spanish-English (267)	0.40	0.47	0.07	0.47 (0.04) 0.35-0.56		
Quechua-Spanish (1032)	0.19	0.17	0.02	0.18 (0.03) 0.10-0.28	r=0.64 , p=0.24	0.91, 0.54-1
Quechua-Spanish (1060)	0.12	0.07	0.04*	0.07 (0.02) 0.03-0.14		
Quechua-Spanish (1075)	0.12	0.14	0.02	0.14 (0.03) 0.06-0.21		
Quechua-Spanish (1077)	0.10	0.11	0.01	0.11 (0.02) 0.06-0.19		
Quechua-Spanish (1081)	0.13	0.05	0.07*	0.05 (0.02) 0.01-0.10		

*Calculated before rounding.

Overall, the strong correlations and small absolute differences observed between the two annotation methods suggest that random sampling is an efficient method for obtaining accurate estimates of children's exposure to different languages and speech registers.

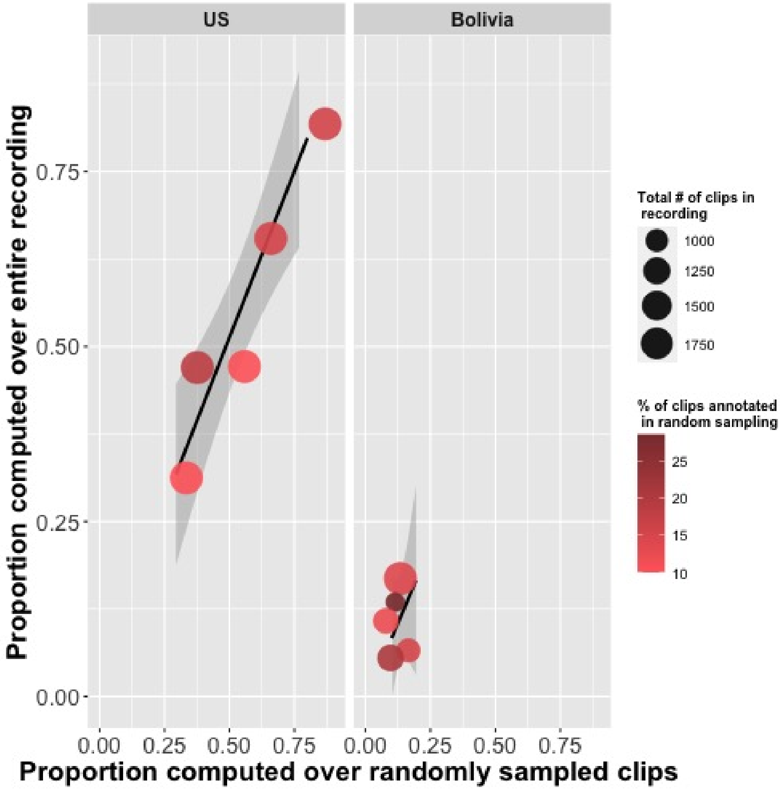


Figure 11. Proportion of child-directed speech clips in U.S. and Bolivian corpora.

5.3 Simulations of random sampling estimates from all-day annotations

In the section above, the random sampling estimates were based on annotations that were conducted prior to the all-day sampling annotation. This was done intentionally to resemble the real-world conditions where annotators would not have listened to the entire recording prior to annotating via random sampling. Since the same coders conducted these annotations at two different times, first for random sampling and then for all-day sampling, imperfect intra-rater agreement likely contributed to the different estimates derived from the two methods: differences in the estimates from random and all-day sampling could be due to differences in the number of clips annotated as well as differences in the judgments made by the annotators at two different times.

To supplement these analyses, we conducted a series of Monte Carlo simulations where we drew 100 random samples from each child's all-day annotations. The goal of these simulations is twofold. First, because the samples were drawn from the all-day annotations, any differences between the simulated random sample estimates and the all-day sampling estimates are due only to differences in the samples themselves (eliminating any differences due to intra-rater agreement). Second, these simulations allow us to obtain a distribution of estimates based on different random samples. The distributions tell us how much estimates of dual language and CDS exposure would vary depending on the random sample we happen to draw. The narrower the distributions, the more likely it is that we would obtain similar results regardless of the sample, thus giving greater confidence that random sampling yields reliable estimates.

We conducted 100 drawings to simulate 100 estimates per child. Each random sample contains the same number of clips used for the original random sampling estimate for that child, as described in section 4.5 (e.g., if 90 clips were originally annotated for a participant during random sampling, we drew 90 clips 100 times for the simulation). See Appendix B for figures displaying the distribution of the simulated estimates of adult-directed and child-directed speech and minority languages. From these simulations, we calculated the mean, standard deviation, and range of the minority language estimate (Table 3) and the CDS estimate (Table 4) for each participant. As expected, the mean of the simulated estimates corresponds well to the estimates based on all-day sampling across all participants. Most importantly, the standard deviation around this estimate is fairly small (from 0.02-0.06 for the minority language estimates and from 0.02-0.06 for the CDS estimates). This suggests that most estimates of minority language and CDS exposure based on random samples are likely to be close to the value based on the entire daylong recording, although they could vary by as much as 0.16.

Finally, in order to evaluate how well different random samples captured variability in children's dual language/CDS exposure, we computed correlations between the corresponding estimates derived from *each* of the simulated random samples and the corresponding estimate derived from all-day sampling. This yielded 100 correlation coefficients for each measure of interest (% Spanish, % Quechua, % CDS); the mean and range of these correlation coefficients are shown in Tables 3 and 4. For the Spanish-English corpus, the correlations for % Spanish between the random sample simulations and the all-day sampling estimates ranged from 0.87-1, with an average of 0.97, suggesting that the random samples did consistently well in capturing variability in infants' exposure to Spanish. The correlations for % Quechua in the Quechua-Spanish corpus ranged from 0.68-1, with an average of 0.96. Although this range is somewhat wider, it again suggests that in most cases random sampling captured variability in infants' exposure to Quechua similar to that obtained from all-day annotation. Finally, and consistent with the results above, the correlations for exposure to CDS in the Spanish-English corpus were consistently strong (M=0.98 [0.89-1]), while those for the Quechua-Spanish corpus were more variable (M=0.91 [0.54-1]).

5.4 Comparisons to parent report

The final research question asked how estimates of bilingual language exposure made via random and all-day sampling compared to those in parental language questionnaires. Detailed parental questionnaires were only collected from the families in the U.S., so this comparison will only be addressed in that corpus. To estimate percent exposure to each language from the parent questionnaires, the number of hours each person was reported to spend with the child *per week* was multiplied by the relative proportion of Spanish (or English) used by that person. This yielded the number of hours per language per person. Then, for each language, number of hours were summed across all people to obtain the number of hours per language per week. Finally, proportion Spanish was computed as the number of Spanish hours divided by total language hours (see Marchman & Martinez-Sussmann 2002 and DeAnda et al. 2016).

Table 5 displays the Spanish language estimates derived from parental questionnaire, as well as both annotation methods (random and all-day). The correlation between Spanish language exposure estimates made from random sampling and parental questionnaire was weak (r(5)=.004, p=.995; Figure 12), as was the correlation between all-day sampling and parental questionnaire (r(5)=.13, p=.84).

Since the parental questionnaire asked caregivers to estimate the amount of Spanish people used when talking *to* the infant, we performed an additional analysis comparing the Spanish language estimates from the questionnaire to the proportion of Spanish in CDS only (Table 5). Limiting the

Table 5

Spanish language estimates in U.S. corpus, by child, register, and estimation method.

	All registers		In child-directed speech		
Child ID	Random	All-day	Random	All-day	Parental Questionnaire
179	0.57	0.57	0.53	0.52	0.74
198	0.87	0.78	0.78	0.64	0.55
199	0.76	0.70	0.64	0.66	0.95
261	0.69	0.65	0.55	0.48	0.73
267	0.92	0.92	0.82	0.87	0.87

Spanish language estimation to CDS resulted in a closer correspondence between recording estimates and questionnaire estimates for only some participants. The correlation between questionnaire and all-day sampling estimates was improved but remained relatively weak (r(5)=0.39, p=.51). On the other hand, the correlation with random sampling estimates was reduced to almost zero (r(5)=-0.07, p=.91). This suggests that estimates of dual language exposure based on parental questionnaires and daylong recordings may differ in what they are capturing, a point we return to in the Discussion.

6 Discussion

This study sought to validate a novel approach—random sampling from daylong audio recordings—to estimate infants' CDS and dual language exposure. Results showed that 1) estimates of the proportion of each language and speech register stabilized after a limited amount of annotation (approximately 7% of the total recording or 11% of all speech clips in the recording), 2) estimates of language and speech register categories made via random and all-day sampling were overall strongly correlated, particularly for higher-frequency categories, 3) simulated estimates of random sampling were consistently correlated with estimates made from all-day sampling, and 4) estimates of language exposure made via random and all-day sampling annotations were weakly correlated with estimates based on parental report. Taken together, these results suggest that random sampling is a promising approach for estimating the proportion of children's exposure to different languages and speech registers as captured in daylong recordings, though caution should be taken when estimating

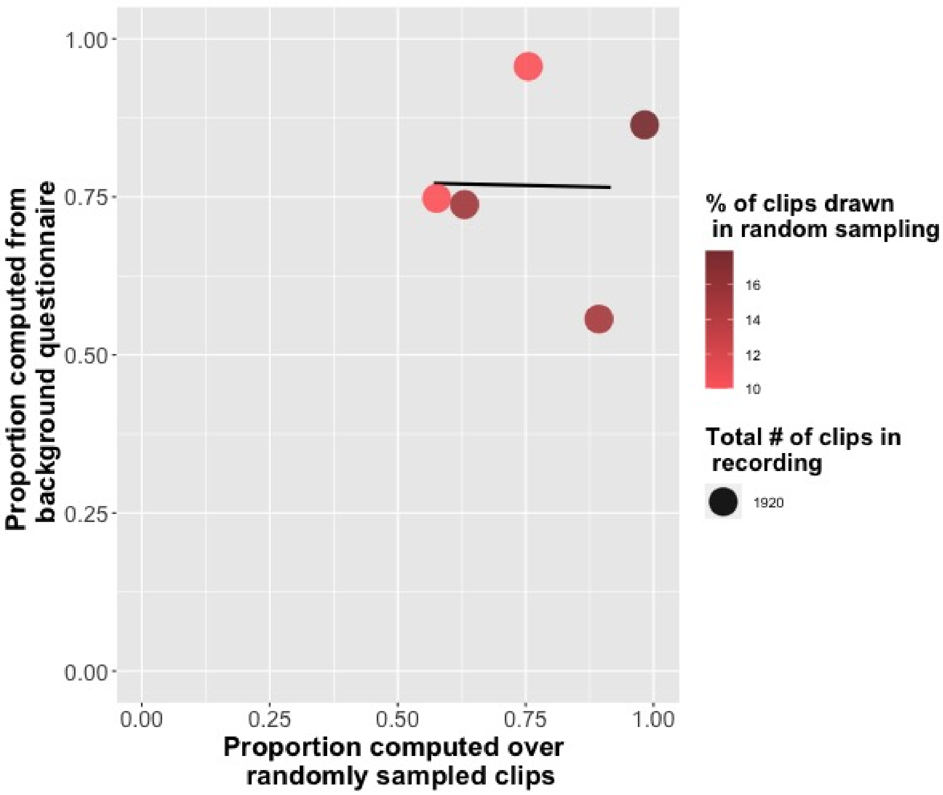


Figure 12. Proportion of Spanish in U.S. corpus: random sampling and background questionnaire methods

infrequent events.

Previous studies that estimated children's exposure to CDS or dual language input required time- and resource-intensive transcription or annotation (Orena et al. 2019; 2020; Weisleder & Fernald 2013) or relied on parental questionnaires (de Houwer 2011; Place & Hoff 2011). Validating estimates of CDS and dual language exposure based on random sampling from daylong audio recordings is an important methodological advance, both because it demonstrates that ecologically-valid estimates of language exposure can be obtained with limited annotation and because relying on naturalistic recordings—instead of written questionnaires—may facilitate the participation of under-served communities in behavioral research.

6.1 Correspondence between CDS and dual language estimates using random and all-day sampling

We found that language and speech register estimates derived from random annotation of daylong recordings corresponded well to estimates derived from all-day annotation, both in absolute value and in capturing variability across children. One exception to this was the relatively weaker, though still moderate (r(5)=.64; p=.24), correlation between CDS estimates made using random and all-day annotation methods in the Quechua-Spanish corpus. This weaker correlation could be due to the limited range of CDS proportions and/or the infrequency of CDS in the Quechua-Spanish recordings from Bolivia. CDS occurred infrequently in this corpus, accounting for, on average, 10.80% (*SD*=4.92) of the overall speech from adults and older children (average determined using all-day sampling). A previous study found that, when the behavior of interest had very low base rates, most sampling strategies provided biased estimates of the behavior (Micheletti et al. 2020). In the current study, we found that the most frequent speech register category in the Quechua-Spanish corpus, adult-directed speech, showed stronger correlations between random and all-day annotation methods (r(5)=.84; p=.08) than did CDS, lending support to the idea that infrequent events may be harder to estimate. On the other hand, the absolute error in the CDS estimates derived from random and all-day sampling for the Quechua-Spanish corpus was fairly low, and similar to that for the Spanish-English corpus. This suggests that the random sampling and all-day sampling estimates were similar, but that restricted range in CDS in the Quechua-Spanish corpus prevented us from observing stronger correlations.

The correspondence observed between the random and all-day sampling estimates is especially impressive given that random sampling annotation was conducted under realistic conditions in which annotators had not yet listened to the entire recording. This demonstrates that it is possible to obtain reliable estimates of CDS and dual language exposure based on samples that are fairly de-contextualized. On the other hand, results from the random sampling simulations suggest that, in some cases, estimates derived from the simulated random samples differed from those based on the original random samples (see Appendix B), suggesting that the additional context gained during all-day annotation may have impacted annotators' judgments. Most importantly, these simulations also show that the correlations between all-day sampling and estimates derived from *different* random samples were consistently strong, suggesting that results from random sampling would be replicable.

Overall, these results suggest that random sampling is an effective method for estimating the frequency of speech registers and dual language input in daylong recordings. However, researchers who suspect that a behavior occurs infrequently in a particular context, or who discover this over the course of annotation, may need to annotate a larger fraction of the recording in order to ensure accurate estimation. Alternatively, researchers could consider annotating fewer levels of a category in order to avoid categories with very low frequencies. For example, the categories target child-directed speech and other child-directed speech could be combined into one CDS category. Finally, estimating low frequency categories may require not only sampling a larger number of segments from a daylong recording, but also collecting a larger number of recordings (e.g., over multiple days) in order to ensure that sufficient examples are captured.

6.2 Correspondence between dual language exposure estimated from recordings and parent questionnaires

Estimates of infants' exposure to Spanish based on recordings from the U.S. corpus did not correspond well to estimates based on parental report. This could indicate that the parental questionnaire responses did not accurately reflect the infants' language exposure or, alternatively, it could indicate that a single daylong recording was not sufficient to estimate the infants' dual language exposure more broadly. The Bilingual Background Interview is intended to capture information about a child's dual language exposure during a typical week. Thus, it is possible that the parental questionnaires captured a wider range of speakers and interactions than did a single daylong recording, offering a better representation of the child's dual language exposure. Indeed, although all of the mothers from the U.S. corpus considered themselves the infant's primary caregiver, several of them reported differences in infants' caregiving environments between weekdays and weekends. In particular, infants tended to spend more time with fathers on weekends than on weekdays, and often also spent time with grandparents and extended family. Since the recordings were conducted on a single day, differences in language exposure between weekends and weekdays were not captured in estimates derived from the recordings.

On the other hand, there are reasons to believe the parent reports may not have accurately reflected children's dual language exposure. When responding to the interviews, mothers often expressed uncertainty about the degree to which different family members used English and Spanish with the child. The majority of the mothers in this study were Spanish-dominant, and several had limited proficiency in English. It may thus have been difficult for them to estimate their infants' English exposure, especially if they did not participate in those interactions. Indeed, as shown in Table 5, the majority of mothers over-estimated the proportion of Spanish children heard relative to the daylong recordings. Asking a single caregiver to report on the language the child hears when they are spending time with other family members may not always provide an accurate representation of these interactions. For example, older siblings/children in the U.S. corpus may have more balanced or English language dominance than their adult caregivers.

Finally, it is important to note that the parent questionnaires and recordings were not collected concurrently. Four of the parent interviews were conducted when infants were 2 months and one when the infant was 28 months, while the recordings included in the current analyses were collected when infants were between 6 and 12 months.[2] Changes in infants' environments during this time period may have contributed to weaker associations between the estimates derived from the recordings and the parental questionnaires. Indeed, several mothers reported they were not working at the time of the initial interview, but were expecting to return to work in the near future. This may have resulted in changes in families' caregiving arrangements that could have significantly impacted infants' dual language exposure between the time of the interview and the recording. Future work should consider carrying out parental language exposure questionnaires concurrently with data collection.

Previous work comparing dual language exposure estimates from naturalistic recordings and parental questionnaire has been mixed. Orena et al. (2020) found that parental questionnaires mirrored language exposure estimates from all-day recording annotations; however, that study estimated language exposure over three daylong recordings, including weekdays and weekends. Crucially, the study also found variability in proportional exposure to each language by recording day, suggesting that a single daylong recording may not be representative of a child's dual language exposure over longer periods. Elsewhere, dual language estimates from daylong recordings better predicted 3-year-olds' speech-language outcomes than estimates from parental report (Marchman et al. 2017). However, in that study, estimates of dual language exposure based on the daylong recording included a measure of the *number of words* in each language, while the estimates based on parental report included only a measure of the *proportion* of each language. Consequently, the naturalistic assessment could be the stronger predictor because, unlike the parental questionnaire, it included variability in the absolute amount of speech heard in each language, rather than just the relative amount.

The current study will not definitively conclude which method of estimating infants' dual language exposure—parental questionnaire or naturalistic assessment—is most accurate. However, it is important to note that both Marchman et al. (2017) and the current work studied bilingual Spanish-English children and infants in the United States (the Bay Area and New York City, respectively), whereas Orena et al. (2020) studied bilingual French-English infants in Montreal. The sociolinguistic contexts in the two locations are highly distinct, with likely implications for dual language self-report. Bilingualism and French language competence are highly valued, and expected, in Montreal and throughout Quebec. In contrast, Spanish-English bilingualism does not enjoy the same prestige in the U.S., even in linguistically-diverse areas such as New York City. Indeed, Orena et al. (2020) note that language and bilingualism are prominent topics in Montreal, which may have led the caregivers to pay particular attention to the languages that their infants were exposed to. The authors also suggest that generous parental leave policies in Canada, which allow caregivers to stay at home with infants for longer, could mean that caregivers are more aware of their infants' dual language exposure. In contrast, the participants in the Spanish-English corpus were all first-generation immigrants to the U.S., and likely faced obstacles and potentially discrimination for speaking Spanish. It is conceivable that, in contexts in which minority languages are routinely devalued, caregivers may not find it easy to report on their child's exposure to the minority language.

Future work should evaluate how dual language exposure (and other parameters like speech register) derived from random sampling vary over multiple recording days to determine if the weak correlations between parental report estimates and naturalistic estimates are due to the limitations of a single recording or instead indicate that parental reports do not accurately capture dual language exposure in certain contexts. Researchers may also consider that, although these methods are not interchangeable, each provides useful and complementary information for understanding children's dual language exposure. Parental questionnaires may be able to provide a more complete picture of the different people children interact with and the languages they use, while daylong recordings may provide more accurate and precise estimates of the language children hear in those interactions.

6.3 Limitations

The estimates used in this study focused on the proportions of language and speech register categories in daylong recordings, not the absolute quantities of each measure. Critically, Orena et al. (2020) demonstrated that the raw *amount* of language exposure can differ between infants exposed to similar language *proportions*, and several works have found that absolute measures of exposure, like the number of words in each language that bilingual children are exposed to, are better predictors than proportional measures (de Houwer 2011, Marchman et al. 2017). In the future, proportional estimates derived from random sampling could be combined with estimates of absolute quantities of speech to better explain individual differences between children.

Although data samples from daylong recordings are highly naturalistic, and arguably more representative than shorter recordings or recordings made in the lab, even a 16-hour at-home recording does not capture all the complexities of a child's language learning environment. Here, we computed dual language and CDS exposure from a single daylong recording for each infant. It is unclear if or how these estimates differ from one day of recording to the next. Some work suggests that there is high variability between daylong recordings collected over multiple days. Anderson & Fausey (2019) compared environmental language measures taken from multiple days of infants' (0;6-1;0) daylong recordings and found that the number of words spoken by adults varied by day. Orena et al. (2020) came to a similar conclusion for infants' dual language exposure. Thus, although our results suggest that estimates based on random sampling provide an accurate representation of the proportion of CDS/minority language exposure in a single daylong recording, they do not necessarily provide a representative estimate of exposure in children's lives. Future research could evaluate the extent of variability in dual language and CDS exposure over multiple timescales, hour-to-hour (within a day), day-to-day (within a week), week-to-week (within a month), and month-to-month. Random sampling from the recordings, which significantly decreases the time required for annotation, should facilitate transcription of these additional recordings.

Finally, this work studied CDS and dual language exposure in two bilingual communities with distinct sociolinguistic profiles. For example, most caregivers in the Quechua-Spanish corpus were balanced bilinguals while the caregivers in the Spanish-English corpus spoke English as a second language. CDS was also notably less frequent in the Quechua-Spanish corpus than the Spanish-English corpus. Random sampling appeared to be relatively robust to these differences across communities. Nevertheless, the results may not apply to all multilingual settings. In communities with more widespread code-switching, monolingual clips may be less frequent, making it more difficult to estimate the proportion of languages in the recording. Trilingual exposure would also decrease the frequency of individual language categories in the recording, possible making it difficult to reliably estimate exposure on the basis of random sampling due to the reduced frequency of each language.

7 Conclusion

The language that children hear in their everyday environments predicts later speech-language outcomes. Daylong audio recordings are a valuable methodological tool to capture children's language experiences, but detailed annotation is time- and resource-intensive. This study validated a novel approach—random sampling from daylong recordings—to estimate children's dual language and child-directed speech exposure in bilingual communities in Bolivia and the United States. By sampling randomly from daylong audio recordings, we demonstrated that stable estimates of infants' exposure to speech register (target child-directed speech, adult-directed speech, other child-directed speech) and language (Quechua or Spanish) were achieved after annotating approximately 7% of the entire recording, or 11% of all speech clips in the recording. Estimates from random sampling were, in general, comparable to those made over the entire recording. This method worked well in both corpora, despite sociocultural differences such as the frequency of certain language and speech register categories; however, results were less clear when estimating events that occurred infrequently and had limited variability across participants. In addition, estimates derived from the daylong recordings did not coincide with those from parental questionnaires collected in the U.S. corpus, though it was not clear if this was due to inaccuracies in parents' report, differences of when the parental questionnaire was administered to the date of the daylong recording, or to deriving estimates from a single daylong recording. Going forward, random sampling of daylong recordings may be an efficient, ecologically-valid method for quantifying children's naturalistic speech-language exposure, though researchers may want to collect recordings over multiple days or contexts to obtain representative estimates of children's ambient language exposure.

Acknowledgements

The authors thank the families who participated in this research program and acknowledge funding from the following sources: National Institute on Deafness and Other Communication Disorders grants T32DC000046 (M.C.) and R21DC018357 (A.W.), Oswalt Documenting Endangered Language Grant (M.C.), Academic Pediatric Association Young Investigator Award (A.W.), and

International Congress of Infant Studies Summer Fellowship (A.V.). Additional thanks to Jan Edwards for the use of her recording equipment to construct the Bolivia corpus and to Alan Mendelsohn and the BELLE team for support in creating the U.S. corpus.

Footnotes

[1] See Mehl & Pennebaker (2003) and Mehl (2017) for details on longform recordings in adult populations.

[2] Since one parental questionnaire was completed at 28 months, we calculated an additional correlation, excluding that participant, between Spanish language estimates derived from from the daylong recordings and those from the parental questionnaire. The correlation between Spanish language estimates derived from random sampling and the parental questionnaire remained weak and insignificant (r(5)=-.12, p=.88) as did the correlation between estimates derived from all-day sampling and the parental questionnaire (r(5)=.08, p=.92).

8 References

Anderson, H., & Fausey, C. (2019). *Modeling non-uniformities in infants' everyday speech environments.* Baltimore, MD. (Presentation given at the 2019 Biennial Meeting of the Society for Research in Child Development)

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. doi: 10.1111/desc.12715

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What Do North American Babies Hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724. doi: 10.1111/desc.12724

Bijeljac-Babic, R., Serres, J., Höhle, B., & Nazzi, T. (2012). Effect of bilingualism on lexical stress pattern discrimination in French-learning infants. *PLoS ONE*, *7*(2), e30843. doi: 10.1371/journal.pone.0030843

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, *16*(1), 32–48. doi: 10.1017/S1366728912000120

Byers-Heinlein, K., Schott, E., Gonzalez-Barrero, A. M., Brouillard, M., Dubé, D., Jardak, A., ... Tamayo, M. P. (2019). MAPLE: A Multilingual Approach to Parent Language Estimates. *Bilingualism: Language and Cognition*, 1–7. doi: 10.1017/S1366728919000282

Carbajal, M. J., & Peperkamp, S. (2019). Dual language input and the impact of language separation on early lexical development. *Infancy*, *25*(1), 22–45. doi: 10.1111/infa.12315

Carroll, S. E. (2017). Exposure and input in bilingual development. *Bilingualism: Language and Cognition*, *20*(1), 3–16. doi: 10.1017/S1366728915000863

Casillas, M., Brown, P., & Levinson, S. (2019). Early language experience in a Tseltal Mayan village. *Child Development*, *0*(0), 1–17.

Casillas, M., Brown, P., & Levinson, S. (2020). Early language experience in a Papuan community. *Journal of Child Language*, *0*(0), 1–23.

Cristià, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2017). Child-Directed Speech Is Infrequent in a Forager-Farmer Population: A Time Allocation Study. *Child Development*, *90*(3), 759–773. doi: 10.1111/cdev.12974

Cychosz, M. (2018). *Cychosz HomeBank Corpus.* doi: 10.21415/YFYW-HE74

Cychosz, M. (2020). *Phonetic development in an agglutinating language.* Berkeley, CA: University of California, Berkeley. (Unpublished doctoral dissertation)

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., … Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods*, 1–19. doi: 10.3758/s13428-020-01365-9

de Bree, E., Verhagen, J., Kerkhoff, A., Doedens, W., & Unsworth, S. (2017). Language learning from inconsistent input: Bilingual and monolingual toddlers compared: Language learning from inconsistent input. *Infant and Child Development*, *26*(4), e1996. doi: 10.1002/icd.1996

de Houwer, A. (2011). Language input environments and language development in bilingual acquisition. In L. Wei (Ed.), *Applied Lingusitics Review* (pp. 221–240). Berlin/New York: De Gruyter Mouton.

DeAnda, S., Bosch, L., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2016). The language exposure assessment tool: Quantifying language exposure in infants and children. *Journal of Speech Language and Hearing Research*, *59*(6), 1346–1356.

Ganek, H., Smyth, R., Nixon, S., & Eriks-Brophy, A. (2018, September). Using the Language ENvironment Analysis (LENA) System to Investigate Cultural Differences in Conversational Turn Count. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2246–2258. doi: 10.1044/2018_JSLHR-L-17-0370

Gilkerson, J., & Richards, J. (2009). *Impact of Adult Talk, Conversational Turns, and TV During the Critical 0-4 Years of Child Development* (2nd Edition ed.) (No. ITR-01-2).

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing Children's Home Language Environments Using Automatic Speech Recognition Technology. *Communication Disorders Quarterly*, *32*(2), 83–92. doi: 10.1177/1525740110367826

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H. Brookes Publishing.

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., … Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science*, *26*(7), 1071–1083. doi: 10.1177/0956797615581493

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5), 1368–1378. doi: 10.1111/1467-8624.00612

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, *11*(6), F31-F39. doi: 10.1111/j.1467-7687.2008.00768.x

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in linear mixed-effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Mahr, T., & Edwards, J. (2018). Using language input and lexical processing to predict vocabulary size. *Developmental Science*, *21*(6), 1–14. doi: 10.1111/desc.12685

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: Comparing direct observation and parent-report measures of dual-language exposure. *Developmental Science*, *20*(1), e12425. doi: 10.1111/desc.12425

Marchman, V. A., & Martinez-Sussmann, C. (2002). Concurrent Validity of Caregiver/Parent Report Measures of Language for Children Who Are Learning Both English and Spanish. *Journal of Speech, Language, and Hearing Research*, *45*, 983–97. doi: 10.1044/1092-4388(2002/080)

Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A Method for the Naturalistic Observation of Daily Social Behavior. *Current Directions in Psychological Science*, *26*(2), 184–190. doi: 10.1177/0963721416680611

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857–870. doi: 10.1037/0022-3514.84.4.857

Micheletti, M., de Barbaro, K., Fellows, M. D., Hixon, J. G., Slatcher, R. B., & Pennebaker, J. W. (2020). Optimal sampling strategies for characterizing behavior and affect from ambulatory audio recordings. *Journal of Family Psychology*. doi: 10.1037/fam0000654

Newman, R. S., Rowe, M. L., & Bernstein Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. doi: 10.1017/S0305000915000446

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2019). Reliability of the Language Environment Analysis (LENA) in French-English Bilingual Speech. *Journal of Speech Language and Hearing Research*, *67*(2), 2491–2500. doi: 10.31234/osf.io/3xcvu

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2020). What do bilingual infants actually hear? Evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental Science*, *23*(2), 1–14. doi: 10.1111/desc.12901

Pearson, B. Z., Fernandez, S. C., Lewedeg, V., & Oller, D. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics*, *18*(1), 41–58. doi: 10.1017/S0142716400009863

Place, S., & Hoff, E. (2011). Properties of Dual Language Exposure That Influence 2-Year-Olds' Bilingual Proficiency: Dual Language Exposure and Bilingual Proficiency. *Child Development*, *82*(6), 1834–1849. doi: 10.1111/j.1467-8624.2011.01660.x

Place, S., & Hoff, E. (2016). Effects and noneffects of input in bilingual environments on dual language skills in 2 $\frac{1}{2}$-year-olds. *Bilingualism: Language and Cognition*, *19*(5), 1023–1041. doi: 10.1017/S1366728915000322

Potter, C. E., Fourakis, E., Morin-Lessard, E., Byers-Heinlein, K., & Lew-Williams, C. (2019). Bilingual toddlers' comprehension of mixed sentences is asymmetrical across their two languages. *Developmental Science*, *22*(4), e12794. doi: 10.1111/desc.12794

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. doi: 10.1111/desc.12172

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017). The Impact of Early Social Interactions on Later Language Development in Spanish-English Bilingual Infants. *Child Development*, *88*(4), 1216–1234. doi: 10.1111/cdev.12648

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science*, *29*(5), 700–710. doi: 10.1177/0956797617742725

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development, and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205.

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development: Child-Directed Speech and Vocabulary. *Child Development*, *83*(5), 1762–1774. doi: 10.1111/j.1467-8624.2012.01805.x

RStudio Team. (2020). *RStudio: Integrated Development for R.* Boston, MA: RStudio, Inc.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, *15*(4), 426–445. doi: 10.1177/1367006911403202

Unsworth, S., Chondrogianni, V., & Skarabela, B. (2018). Experiential Measures Can Be Used as a Proxy for Language Dominance in Bilingual Language Acquisition Research. *Frontiers in Psychology*, *9*(1809), 1–15. doi: 10.3389/fpsyg.2018.01809

Usoltsev, A. (2015). *Voice Activity Detector-Python.* Retrieved from `https://github.com/marsbroshok/VAD-python` (GitHub Repository)

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152. doi: 10.1177/0956797613488145

Weisleder, A., & Mendelsohn, A. (2019). *Daylong recordings of 2-12 month-old infants from spanish-speaking homes in the u.s.*

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag New York.

Appendix A

Table A1
Number of clips annotated by child and annotation method.

	# of clips annotated (% of total clips)	
Corpus (ID)	Random	All-day
Spanish-English (267)	101 (5.26 %)	274 (14.27 %)
Spanish-English (261)	92 (4.79 %)	294 (15.31 %)
Spanish-English (199)	118 (6.15 %)	467 (24.32 %)
Spanish-English (198)	81 (4.22 %)	302 (15.73 %)
Spanish-English (179)	120 (6.25 %)	633 (32.97 %)
Quechua-Spanish (1081)	92 (7.5 %)	285 (23.25 %)
Quechua-Spanish (1077)	83 (7.23 %)	355 (30.92 %)
Quechua-Spanish (1075)	81 (8.69 %)	199 (21.35 %)
Quechua-Spanish (1060)	111 (10.51 %)	405 (38.35 %)
Quechua-Spanish (1032)	97 (5.05 %)	372 (19.38 %)

Appendix B

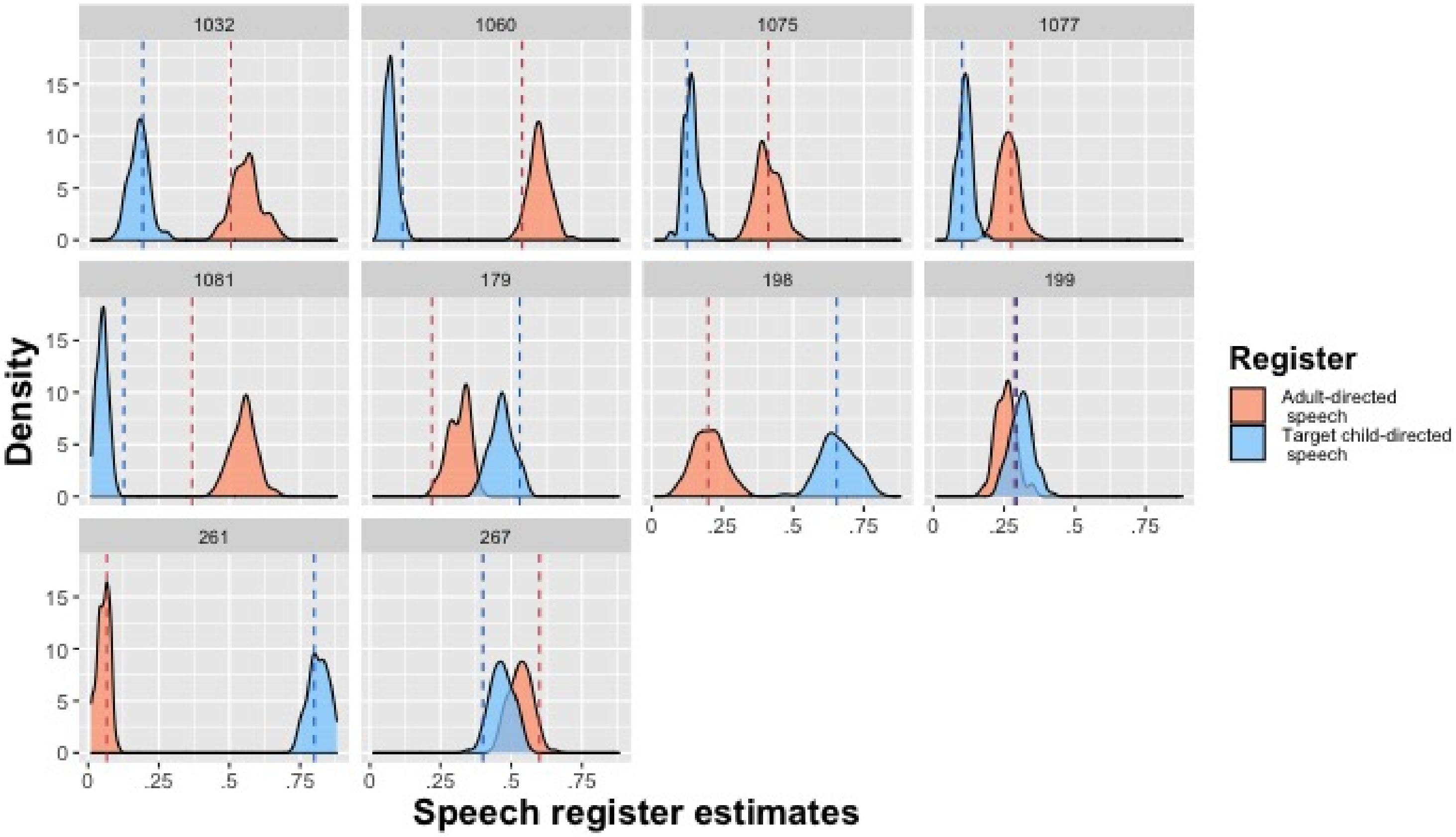


Figure B1. Estimates of adult- and child-directed speech quantities, by child. Density curves indicate the distribution from n=100 simulated estimates. Overlapping distributions indicate similar quantities of adult- and child-directed speech in the child's environment. Dotted lines indicate the estimate derived from the original random sampling method.

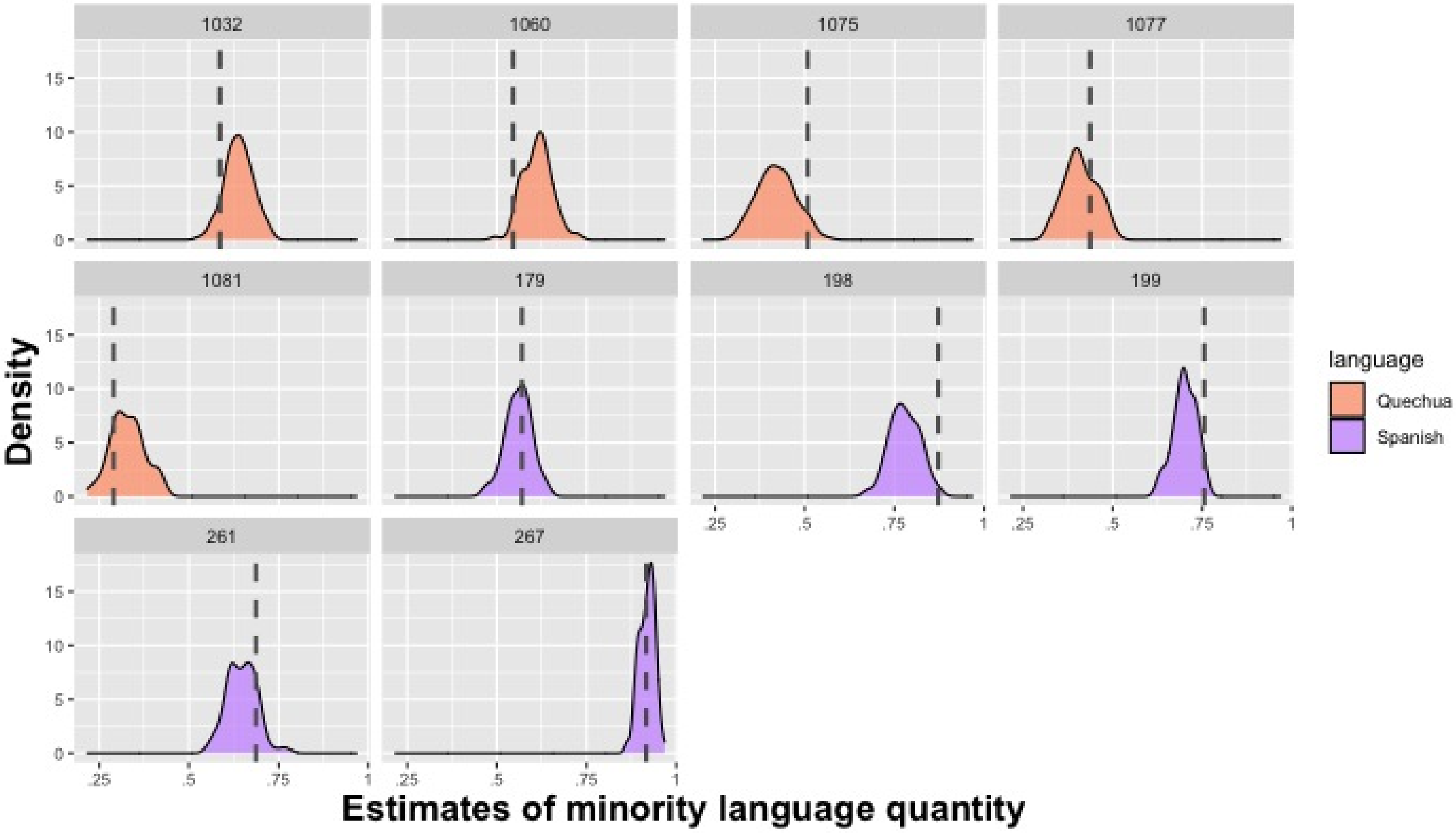


Figure B2. Estimates of minority language quantities, by child. Density curves indicate the distribution from n=100 simulated estimates. Dotted lines indicate the estimate derived from the original random sampling method.