Establishing the reliability of metrics extracted from long-form recordings using LENA and the ACLEW pipeline

Alejandrina Cristia[1], Lucas Gautheron[1,2], Zixing Zhang[3], Björn Schuller[4,12], Camila Scaff[1,5], Caroline Rowland[6], Okko Räsänen[7], Loann Peurey[1], Marvin Lavechin[1], William Havard[8], Caitlin Fausey[9], Margaret Cychosz[10], Elika Bergelson[11], Heather Anderson[9], Najla Al Futaisi[12], Melanie Soderstrom[13]

[1] Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, France

[2] University of Wuppertal, Interdisciplinary Centre for Science and Technology Studies (IZWT) Wuppertal, Nordrhein-Westfalen, Germany

[3] Hunan University, School of Computer Science and Electronic Engineering Changsha, Hunan, China

[4] Technische Universität München, MRI, CHI - Chair of Health Informatics, Munich, Germany

[5] University of Zurich, Human Ecology group, Institute of Evolutionary Medicine Zurich, Switzerland

[6] Max Planck Institute for Psycholinguistics,Nijmegen, The Netherlands

[7] Tampere University, Unit of Computing Sciences, Unit of Computing Sciences Tampere, Finlande

[8] Université d'Orléans, Centre-Val de Loire, France

[9] University of Oregon, Eugene, OR, USA

[10] University of Maryland at College Park, Department of Hearing and Speech Sciences College Park, MD, USA

[11] Harvard University, Psychology, Cambridge, MA, USA

[12] Imperial College London, GLAM – Group on Language, Audio, & Music,
London, UK
[13] University of Manitoba, Psychology, Winnipeg, MB USA

Author note

This work was funded by: Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex); the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095); Academy of Finland (grant no. 314602); Social Sciences and Humanities Research Council of Canada, 869-2016-0003 (ACLEW) and 435-2015-0628; NEH HJ-253479-17; NIH DP5 OD019812-01. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) and the Bridges system funded by National Science Foundation (OCI-1053575, ACI-1445606, Pittsburgh Supercomputing Center (PSC)); CSC – IT Center for Science, Finland; GENCI-IDRIS, France (Grant-A0071011046). We thank Nicolas Rochat for input at the beginning of this project. Authors are sorted reverse-alphabetically, except for the first, second, and last.

Abstract

Long-form audio recordings are increasingly used to study individual variation, group differences, and many other topics in theoretical and applied fields of developmental science, particularly for the description of children's language input (typically speech from adults) and children's language output (ranging from babble to sentences). The proprietary LENA software has been available for over a decade, and with it, users have come to rely on derived metrics like adult word count (AWC) and child vocalization counts (CVC), which have also more recently been derived using an open-source alternative, the ACLEW pipeline. Yet, there is relatively little work assessing the reliability of long-form metrics in terms of the stability of individual differences across time. Filling this gap, we analyzed eight spoken-language datasets: four from North American English-learning infants, and one each from British English-, French-, American English-Spanish, and Quechua-Spanish-learning infants. The audio data were analyzed using two types of processing software: LENA and the ACLEW open-source pipeline. When all corpora were included, we found relatively low to moderate reliability (across multiple recordings, intraclass correlation coefficient attributed to the child identity (Child ICC), was <50% for most metrics). There were few differences between the two pipelines. Exploratory analyses suggested some differences as a function of child age and corpora. These findings suggest that, while reliability is likely sufficient for various group-level analyses, caution is needed when using either LENA or ACLEW tools to study individual variation. We also encourage improvement of extant tools, specifically targeting accurate measurement of individual variation.

Keywords: daylong recordings; big data; speech technology; accuracy

Introduction

Long-form audio recordings of people's everyday experiences have begun to transform the study of individual variation, group differences, and many other topics in theoretical and applied fields of developmental science (Casillas & Cristia, 2019). This is particularly true for the study of child language development, where these recordings provide a unique window into children's real world speech experiences, both with respect to language input (typically speech from adults) and the child's own language production (for recent reviews and meta-analyses, see Cristia et al., 2020; Ganek & Eriks-Brophy, 2018; Wang et al., 2020). Lengthy audio recordings provide practical challenges for traditional methods of analysis in developmental science, which often rely on painstaking human annotation and transcription, which can take 8-20x the length of the audio to transcribe accurately, even by highly trained human coders (Soderstrom et al., 2021). These traditional techniques are much too slow to reap the significant potential of long-form recordings. Tackling this issue, the LENA Foundation developed software for automated speech analyses adapted to long-form recordings (Greenwood et al., 2011). For fifteen years, this software has been the primary source of automated approaches to analyzing these recordings, and providing derived metrics like adult word count (AWC; the number of words produced by adults within the child's close environment); child vocalization counts (CVC; the number of speech-like vocalizations produced by the child); and conversational turn counts (CTC; the number of child-adult or adult-child vocalization sequences).

More recently, open source alternatives to LENA have emerged. Despite the widespread use of derived metrics like AWC, CVC, and CTC, there is relatively little work assessing these and other long-form metrics' reliability as a measure of individual differences (i.e., to what extent individuals' relative scores are stable across repeated measurements). That is, are these metrics like the MacArthur Bates-Communicative Development Inventory

(MB-CDI), another commonly used language metric for toddlers, which shows test-retest correlations >.8 among American participants (Fenson et al., 1994)? Or are they more like lab-based speech perception tasks, that show no test-retest reliability (Cristia et al., 2016)? We aim to document the reliability of metrics derived from LENA and open-source alternatives through corpora collected from linguistically and culturally diverse families, and, as much as possible, benchmark reliability against alternative metrics of children's early language development. In the remainder of this Introduction, we first introduce three key terms: accuracy, reliability, and validity. We then provide a brief overview of extant metrics derived from long-form recordings. We then summarize work on the accuracy, reliability, and validity of long-form-derived metrics as well as potentially comparable language development screeners and tests. We end this Introduction with a brief overview of this study's key contributions.

Assessing Accuracy and Reliability

Probing the strengths and limitations of LENA’s metrics, as well as newer alternatives, is crucial given the growing importance and influence of these systems in the literature (e.g., Ganek & Eriks-Brophy, 2018). One important consideration is *accuracy* of the metric, i.e. the extent to which automatically derived metrics reflect reality (typically benchmarked by what trained human annotators would indicate for the same measure or construct). For instance, accuracy would assess whether when the system attributes a section of the audio to the target child (CHN in the LENA system), this is actually true. Evidence has accumulated on this topic, and has been meta-analyzed in Cristia, Bulgarelli and Bergelson (2020). There have also been some reports on *validity,* which refers to the extent to which a metric accurately represents an intended construct, for instance, whether “child vocalization count” validly represents children's language abilities. A meta-analysis of the validity of

LENA metrics, based on correlations between them and other measures of children's language, can be found in Wang et al. (2020). In this paper, we focus instead on the extent to which a metric is *reliable,* defined as whether individuals' relative scores are stable across repeated measurements.

To understand the significance of reliability, it is helpful to consider how the metrics generated by automated algorithms are often used. Typically, researchers are interested in examining the influence of group or individual differences in these metrics, the relationship of these metrics to each other, or the relationship between these metrics and other measures of child development. These measurements and relationships may be examined concurrently or longitudinally (Wang et al., 2020). For example, researchers may compare the number of words (AWC) spoken around infants from different populations, under the assumption that the measure AWC represents the construct of *children's spoken language input*. AWC might then be used as a concurrent predictor of child vocalization counts (CVC) which are assumed to be related to the construct of *child language skills*. Alternatively, researchers could test the concurrent predictive value of AWC on a totally external measure also thought to capture a construct of child language skills, e.g. the MB-CDI, a parental report of child vocabulary acquisition (Fenson et al., 1994). These comparisons all rely on the assumption that these metrics capture variance related to *trait variation* at the level of the individual (i.e., AWC represents the construct of *input*, CVC and MB-CDI capture the construct of *child language skills*). A strong example of such work is a recent study assessing the construct and criterion validity of LENA metrics among 32 infants learning Hebrew or Arabic (Levin-Asher et al., 2023).

However, like any behavioral measure, metrics derived from long-form recordings may vary for reasons unrelated to trait-based variation. First, high levels of variability could simply be due to the noisiness of the metric itself, rather than meaningful individual variation.

This would be the case if, for example, the AWC metric was not high in accuracy. However, we already know from previous work that automated AWC metrics are highly correlated with human counts of words on the same clips (Cristia et al., 2020). Second, high levels of variability could also be due to *state variation* within and across individuals; for instance, if certain recordings happen to occur when the child is sick or fussy, leading to high incidence of crying, while other recordings happen to contain a great deal of book reading, which are likely to have higher AWC. In such cases, we could find low reliability even if the accuracy (as measured against a human gold standard) is high. Notice that if reliability is low, a substantial portion of the variance observed in a given metric is *not* due to trait-based individual differences. Thus, reliability of a metric constitutes a threat to interpretation not only of first-order studies (comparing individuals) but also to second-order studies (where those same data points are used as predictors of some other measure, concurrently or longitudinally). The methodological quality of a measure is equally important for theory-building. To begin, if we observe low reliability, this may be an indication that our metric does not reflect the psychological construct we had hoped to capture. Additionally, random variation may lead to low replicability of results, constituting a key challenge to theoretical development (see e.g. Grahek et al., 2021).

Are changes with age a threat to reliability? Not necessarily. We describe reliability as a function of individuals' relative scores, rather than individuals' absolute scores, because it more accurately captures the kinds of analyses reported below. These analyses build on variance attributed to individuals above and beyond any *systematic* changes that affect all children similarly. For example, if we rank a group of children at one month of age based on their birth weight, and they all gain the same weight over 5 months, since the same ranking holds at 5 months we can conclude the measure is reliable, even though all of the children’s weights shifted a great deal over the intervening four months. Thus, what is most relevant is not how much absolute scores change, but the stability in children's *relative* scores.

How should reliability be measured statistically? One very useful way to measure the reliability of a test involves collecting data using the same test (or two equivalent versions of the same test) twice over a short period of time, and then calculating the correlation between these two outcome metrics. Some long-form corpora have more than two observations per child, leaving simple correlations (which rely on paired observations) off the table. In such cases, the recommended alternative to a correlation coefficient is the Intra Class Coefficient or ICC (Koo & Li, 2016), which is defined as the proportion of variance that can be attributed to individual variance (as opposed to random error).[1] Thus defined, ICC can be calculated from a mixed linear regression model, which can accommodate more than two data points per individual.

Introduction to long-form recordings

Long-form recordings are typically collected by dressing an infant or child in a custom-designed t-shirt or vest equipped with a breast pocket that exactly fits the recording device, which often runs for many hours (Casillas & Cristia, 2019; Pisani et al., 2021). Given their length, manual annotation of the resulting sound files is prohibitive, so automated analyses are common. We provide here a brief overview of the two main extant processing pipelines, LENA (Greenwood et al., 2011) and ACLEW (Schuller et al., 2024). More detailed descriptions and citations are provided in our Methods and in previous work introducing the pipelines cited therein.

[1] We wanted to provide readers with an intuitive way to interpret ICC values. To this end, we simulated data using correlations between paired data points between .1 and .9, using the same child and corpus distribution as in the studies below, and further constraining our simulated metrics to have the means and standard deviations that they exhibit in the real data. As detailed in SM A, the simulated ICCs provide reasonable approximations of the underlying r's (though they undershoot a bit, e.g., simulated ICCs of .3-.5 were observed for an underlying r = .5). Based on this, we propose that ICC values can be seen as conservatively analogous to correlation values.

The first step of analysis involves “voice type diarization”, whereby sections of the audio containing human vocalizations are detected and attributed to the following broad class talker types: female adults, male adults, key child (wearing the device), and other child. LENA also diarizes other types of audio sections that will not be studied here, namely TV, noise, and overlap (when any of the previous categories overlap with each other). Sections attributed to the key child contain within them sections that are speech-like and sections that are classified as cry or other fixed vocalizations, in addition to short silence between the speech-like, cry, and fixed vocalizations. As mentioned above, this allows the estimation of LENA’s 3 key metrics: AWC, CVC, and CTC.

The ACLEW pipeline was developed years after LENA's (as an open-source alternative), and was conceptually inspired by it, resulting in roughly the same broad processing stages: voice type diarization to find sections of the audio that are attributed to the key child, other children, female adults, and male adults; further analyses of key child vocalizations to distinguish among speech-like, and non-speech-like vocalizations; and further analyses of adult vocalizations to estimate word counts as well as the number of phones and syllables.

In addition to the key metrics, in this work we also derived other metrics from both LENA and ACLEW automated annotations that are less commonly discussed in the literature, such as total vocalization duration and the average duration of a vocalization (see e.g., Cychosz et al., 2023). Some researchers have been employing hourly peak metrics to better capture individual variation (Bergelson et al., 2019; Fibla Reixachs, 2021). That is, rather than using averages across all hours, researchers calculate for example the AWC for each hour in the recording, and find the maximum hourly AWC. Therefore, we additionally included hourly peak values for LENA's CTC and CVC and its ACLEW equivalents. We aimed to provide a rather comprehensive overview of the metrics that could be extracted to summarize vocal activity over a whole day. The full list of possible metrics is shown in Table 1.

Table 1: The 75 metrics studied in the present paper. Number of events and total duration always control for recording length by calculating rates per hour. Pipeline indicates whether a metric is available for both LENA and ACLEW pipelines, or only for one of them. N indicates the number of metrics (e.g., the first line refers to six: number of vocalizations by male adults, female adults, other children, each according to LENA and ACLEW). Number of conversational turns is defined differently across LENA and ACLEW: in LENA, only child vocalizations that are speech-like contribute to it and the gap between adult and child vocalization can be up to 5 seconds, whereas for ACLEW any child vocalization counts but the gap is at most 1 second. "Canonical" are more advanced speech-(like) vocalizations, which could be defined as containing a clear adult-like consonant-vowel or vowel-consonant transition.

Pipeline	Explanation		N
Input metrics (N = 46)			
both	vocalizations by talker type (male adult, female adult, other child)	number	6
		peak hour number	6
		total duration	6
		average duration	6
	words (male adult, female adult, combined)	number (AWC)	6
		peak hour number	6
	conversational turns	number (CTC)	2
		peak hour number	2
ACLEW	syllables (male adult, female adult, combined)	number	3
	phonemes (male adult, female adult, combined)	number	3
Output metrics (key child; N = 29)			

both	crying	number	2
		total duration	2
		average duration	2
	vocalizations (collapsing across types)	number	2
		peak hour number	2
		total duration	2
		average duration	2
	linguistic proportion = (speech)/(cry+speech)	based on number	2
		based on total duration	2
ACLEW	canonical	number	1
		total duration	1
		average duration	1
	non-canonical	number	1
		total duration	1
		average duration	1
	canonical proportion = canonical /(can+noncan)	based on number	1
		based on total duration	1
LENA	speech-like	number (CVC)	1
		peak hour number	1
	Automatic Vocalization Analysis (AVA)		1

Previous work

In this section, we review the psychometric literature on long-form recording metrics before turning to a summary of reliability studies more broadly. Most of the literature assesses *accuracy*, comparing automated metrics against human annotations for the same audio files, meta-analyzed by Cristia et al. (2020). In a nutshell, this previous work finds fairly high correlations between human and automated metrics for LENA's AWC (mean r =

.79, N = 13 studies) and CVC (mean r = .77, N = 5 studies), with lower estimates for CTC (mean r = .36, N = 6 studies). There was insufficient information on more granular estimates of, for example, labels for key child and female adult. A brief discussion regarding which units should and can actually be measured (in a language-independent way) can be found in Räsänen et al. (2021). Although this work establishes accuracy, it is important to remember that high accuracy does not necessarily translate into high reliability.

Additionally, there is a meta-analysis that is potentially relevant: Wang et al. (2020) report medium-sized correlations between automated LENA metrics (AWC, CVC, CTC) and standardized measures of child language (e.g. concurrent or future vocabulary). Does this mean that all three measures are *valid*? One issue is that standardized measures of *child* language may be viewed as tapping the same construct as CVC, but not AWC or CTC. To show that AWC is a good measure of children's input quantities, validity would ideally be assessed against some other measure of e.g. parental volubility. Similarly, construct validity for CTC would have to be established against interactive language, e.g. use of IDS.

There are, to our knowledge, only two peer-reviewed reports focusing on test-retest reliability of LENA metrics (though see Bergelson et al., 2023, supplementary material). Gilkerson et al. (2017) provides one set of reliability estimates, drawn from a larger longitudinal study involving over 300 families of children aged 2-48 months of age, where participants were recorded once per month for up to six months. All participants were thought to be typically developing, and were learning American English monolingually. Between 92 and 248 individual infants provided data at each time point for these particular analyses, each having been recorded several times between one and four months apart. The authors used Pearson pairwise correlations by randomly selecting two data points per child and found moderate correlations for AWC (average r = .44), and higher ones for CVC and CTC (average r = .69 and .71 respectively). In addition, analysis of a randomly-selected sample of

52 participants (whose age range was not reported), who were recorded two days in a row, yielded similar correlations to the ones taken a month apart. Converging results are found in a report on 107 British English monolingual families, whose 24-48 month-old children were recorded on three different days (d'Apice & Stumm, 2019). The authors only reported on AWC, finding correlations of .42-56 across pairs of days, and a global ICC of .47.

In sum, while there is growing attention on the accuracy of long-form recordings, very little work so far addresses their reliability (or their validity). Moreover, previous evaluations have focused on LENA, but there are additional software options now available for developmental scientists (Casillas & Cristia, 2019). These alternatives, such as the open source ACLEW pipeline detailed below, are promising but have not been systematically assessed to the same extent. To begin with, analysis of accuracy comparing automated annotation against human annotation for the ACLEW pipeline has only been reported within the same articles that present the algorithms (Lavechin et al., 2020; Räsänen et al., 2021), raising the possibility of a conflict of interest, in addition to the fact that it has been carried out with much fewer and less diverse samples than research documenting the accuracy of LENA automated annotations. Moreover, there is, to our knowledge, no report of potential convergent validity or reliability for metrics extracted using the ACLEW pipeline. Even within LENA automated annotation, only AWC, CVC, and CTC have been thoroughly assessed, despite the fact that other metrics could be derived from the same data, including total vocalization duration by different talker types and average vocalization duration, and could, in principle, prove to be more reliable than the three most commonly used metrics.

Given that relatively little work has been done on the reliability of long form-derived metrics, we turned to related literature to provide benchmarks against which to interpret the values we come to obtain. We looked at three areas of related literature, focusing on language development since we reasoned that measures of spoken input would be quite different from what we have here (e.g., type-to-token ratios). First, there are alternative measures of language development used with observational data. Second, there are standardized clinical instruments that test language skills. And third, there are studies evaluating the psychometric properties of metrics extracted from wearable data. We discuss each in more detail in Supplementary Materials (SM) B - for the sake of brevity, here we provide a very short discussion of these literatures.

When we looked closely at previous literature, we did not find perfect assessments or measures against which to compare our own results. For instance, measures of language development based on observational data tend to rely on caregiver report, and thus their stability could reflect stability in e.g. parental perception as much as child behavior. One notable exception is a meta-analysis of test-retest in infant laboratory tests, which finds a correlation across days that is not different from zero (Cristia et al., 2016). For standardized clinical instruments, we found only a handful, and in these particular cases, the only reports of reliability we found were by the same companies that sell them, which represents a conflict of interest. An exception is a standardized instrument called Ages and Stages, for which a meta-analysis for 2 to 2.5 year-olds reports r's between .67 and 1 (Velikonja et al., 2017). Relatedly, for older children, a systematic review integrated evidence on test-retest reliability reported by researchers unaffiliated with the test-producing companies (Denman et al., 2017). They found a meta-analytic mean test-retest correlation r=.67, with a range from .35 to .76. Finally, a third meta-analysis focuses on test-retest correlations for metrics extracted from wearables in laboratory conditions, which may overestimate reliability given the controlled environment, with a meta-analysis concluding ICCs are above .6 (Kobsar et al., 2020). Worryingly, the authors of all three meta-analyses just mentioned (Denman et al., 2017; Kobsar et al., 2020; Velikonja et al., 2017) comment on the low quality of studies assessing reliability, strongly suggesting that our field needs to pay particular attention to these issues.

If we collapse across all three meta-analyses and other evaluations (summarized in SM B), we can expect reliability here to be between r = .35 and 1, with levels comparable to meta-analyzed research being at an average r = .6.

Present work

The present study sought to fill a gap in the literature evaluating automated metrics extracted from long-form recordings. We base our analyses on several corpora, collected mainly, but not only, from infants learning American English (which have been the focus of much LENA research). These data were collected independently by several researchers, and are re-used in the context of the present project. Although all recordings were gathered using the LENA hardware, our analyses investigate metrics derived from LENA software analyses as well as an open source alternative developed by the ACLEW project.

Methods

Data

The main characteristics of the eight corpora are summarized in Table 2. Each corpus consists of home recordings of one day (4–16 hours), collected using the LENA recorder, sampling a unique community, with varying dialects across the corpora, and varying socioeconomic statuses both within and between corpora. The data was collected at different sites by different researchers, including two cases of data reuse from prior projects. The largest is the ACLEW project (Soderstrom et al., 2021). Among ACLEW corpora, we focus on four different corpora of child long-form recordings: the Bergelson SEEDLingS corpus ("Bergelson") of American English families from the upstate New York area (Bergelson, 2017), the ESRC LuCiD Centre Language 0-5 corpus ("LuCiD") formed by British English-speaking families living in the North West of England (Rowland et al., 2018), the McDivitt and Winnipeg ("Winnipeg") corpus of families from the Manitoba region who are mostly Canadian English speakers (McDivitt & Soderstrom, 2016), and the Warlaumont ("Warlaumont") corpus of children growing up in Merced, California (Warlaumont et al., 2016), who are mostly American English/Spanish speakers. We additionally included the Cougar corpus[2] ("Cougar") of American English-speaking families based in Washington State (VanDam, 2018); the Fausey-Trio ("Fausey-Trio") corpus of American English-speaking families based in Eugene and Springfield, Oregon (Fausey & Mendoza, 2018); the Lyon corpus ("Lyon") of French-speaking families based in Lyon, France (Canault et al., 2016); and the Quechua corpus ("Quechua") of bilingual Spanish-Quechua speaking families living in rural areas outside of a mid-sized town in the southern Bolivian highlands (Cychosz, 2022). Some Bergelson recordings, and all Winnipeg, Fausey-Trio, Lyon, and Warlaumont recordings are available in the HomeBank repository (VanDam et al., 2016); the LuCiD recordings are available in the Language Archive (Drude et al., 2012). Given our interest in stability across long-form recordings, we excluded recordings shorter than 4 hours, and children who had only contributed one recording.

Table 2: Data distribution within each corpus. Recs/child indicates the number of recordings per child. Recs total is the total number of recordings in the corpus that met inclusion criteria. Duration (h) indicates the average number of hours in each recording. Ages are given in months.

Corpus	Location	Children	Recs/ child	Recs total	Mean Duration (h)	Mean Age	Age Range
bergelson	Northeast US	44	10-12	522	14.0	11.2	6-17
cougar	Northwest US	26	3-45	239	11.1	26.6	0-59
fausey-trio	Western US	28	3	84	13.7	8.9	6-12
lucid	Northwest England	35	4-7	235	15.4	20.0	10-31

[2] This corpus contains a large proportion of recordings of children with hearing loss. The recordings of these children were excluded from the present study.

lyon	Central France	16	3-6	54	14.1	19.6	2-41
quechua	Bolivian Highlands	20	2-3	48	12.5	37.6	9-78
warlaumont	Western US	13	2-4	38	12.5	6.3	2-18
winnipeg	Western Canada	9	2-5	34	9.1	14.5	2-33
Overall		191	2-45	1254	13.5	17.0	0-78

Processing

We processed the long-form recordings using two different pipelines, one of them based wholly on LENA products, and the other with software purpose-built in the ACLEW project. There is no overlap in data used for training the LENA software and the present study, but there is a small level of overlap between the data used to train the ACLEW pipeline, which is so small that we doubt it will affect results.[3]

Given that LENA has been extensively presented in previous work (e.g., Ganek & Eriks-Brophy, 2018; Gilkerson et al., 2017), we do not dwell on its technical characteristics here. The tools that are part of the ACLEW open-source pipeline are relatively new, but have also been introduced in previous work (Al Futaisi et al., 2019; Lavechin et al., 2020; Räsänen et al., 2021). For both pipelines, a similar log is generated, corresponding to the whole audio length with specific sections attributed to four key talker types (key child, other child, male adult, female adult), and with estimated word counts available for female and male adult talkers. There are only three key differences between the logs returned by LENA versus ACLEW tools. First, in LENA, vocalizations from different talker types cannot overlap (such instances are classified as ‘overlap’ and thus excluded from the counts), whereas they can in ACLEW (and are therefore included in counts). Second, adult sections additionally have syllable and phoneme count estimations in ACLEW. Third, each individual key child

[3] Specifically, one of the ACLEW components (which are presented below) used 20 minutes and the two others used 4h of hand-annotated data out of the 11,027h in the 4 ACLEW datasets included here (Räsänen et al., 2021); and one additionally used for 20 out of the 522 Bergelson recordings (Al Futaisi et al., 2019).

vocalizations is classified "canonical", "non-canonical", "crying", or "other" in ACLEW, whereas in LENA, each child vocalization can be split up into a sequence of these subtypes (e.g., the first second is a cry, then an interstitial pause, then a speech-like section[4]).

Analyses

We processed each audio file in our corpora (N=1,254 recordings) through the two pipelines (LENA, ACLEW) to compute the metrics in Table 1.[5] Notice that we have controlled for effects of differences in recording length by using hourly rates for all metrics that are counts or total duration. Controlling for recording length did not seem necessary for linguistic or canonical proportion, nor any of the average duration metrics. For all analyses, we scaled variables by subtracting the mean and dividing by the standard deviation (across all participants and corpora), to more easily compare results across metrics and analyses.

We first performed correlation-based analyses to facilitate comparison with Gilkerson et al. (2017). As we explain below, these overestimate the reliability of the metrics. Therefore, for our main analyses, we instead fit one mixed effects linear regression predicting one metric at a time (e.g., AWC) from child age as fixed effect, with each child nested within corpus as a random effect. We estimate cross-recording reliability by assessing the proportion of variance attributed to the child random effect, that is, the Child ICC using the performance package (Lüdecke et al., 2021). In intuitive terms, Child ICC reveals how correlated repeated measures of the same child are (controlling for age and corpus), and this for each individual input and output metric. A high Child ICC for an input metric means that children who have a higher value in that metric than other children for one recording also tend to have relatively higher value for other recordings; i.e., relative properties of the child's speech input are stable across recordings. A high Child ICC for an output metric entails that e.g. a child that

[4] In the version of the LENA automated analyses software used for analyzing the data in this paper, canonical vocalization counts were not available..

[5] Two recordings associated with a single child in the Winnipeg corpus could not be analyzed with ACLEW due to ethical concerns around data sharing. In addition, one recording from Fausey-trio and one from Quechua could not be included in our LENA metrics because the LENA output file was corrupted or missing.

vocalizes more than others (controlling for age and corpus) at one time point is also a relatively high vocalizer at other time points; i.e., relative properties of the child themself are stable. We thought readers may have a hard time imagining what high or low Child ICC looks like, so we provide an illustration of the level of association across two recordings taken from the same child for the two metrics with extreme Child ICC found in our data in Figure 1. A variable with one of the lowest Child ICC (LENA's peak hourly AWC, Child ICC = .23) has a flatter regression line and wider confidence intervals, with points spread widely around the regression line. In contrast, the variable with the highest Child ICC (ACLEW's other child vocalization duration, Child ICC = .64) has points that are less scattered and closer to the diagonal. This is because when individual variation is very stable, children who had a relatively low number in a given metric in one recording also have a relatively low number in the same metric extracted from the second recording, leading to more points and the regression line being closer to the 45° line.

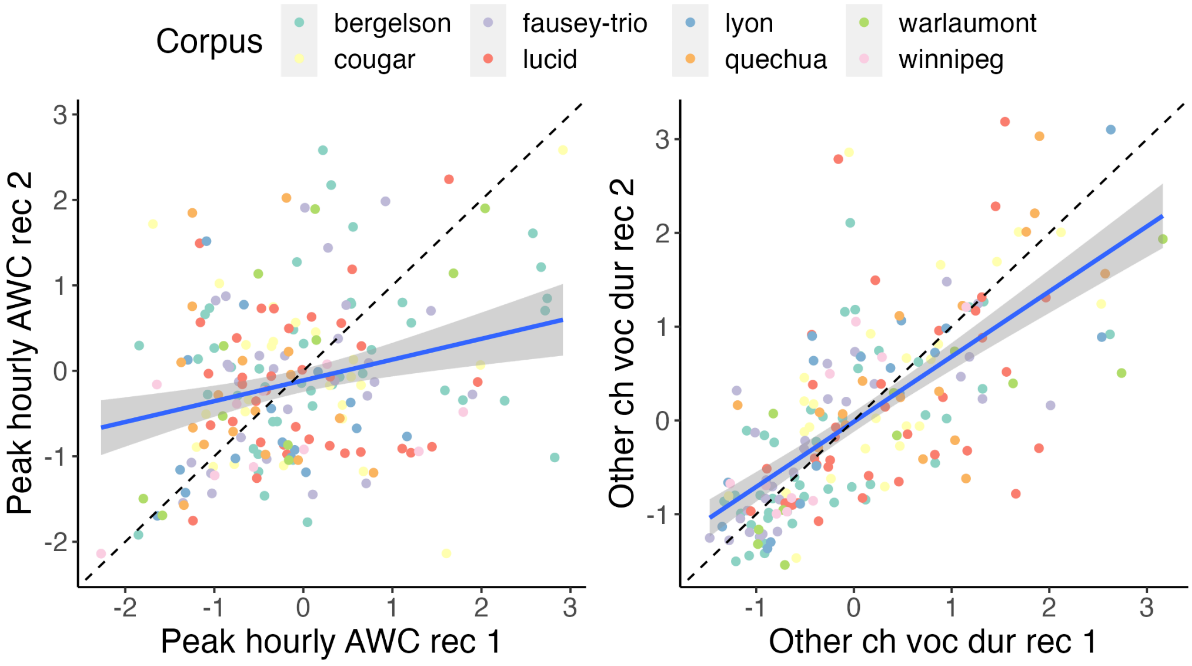

Figure 1. Scatterplots of a given metric across two randomly-selected recordings for each child; for a variable with the lowest Child ICCs (peak AWC per hour, on the left) and the

highest Child ICCs (duration of vocalizations by other children, right). To facilitate inspection, both metrics have been z-scored. Children's individual metrics will be less correlated for variables with low ICC than those with high ICC.

In more technical terms, we used lme4 (Bates et al., 2015) in R (R Core Team, 2023) to fit models of decreasing random effect complexity, as illustrated in Figure 2. We first attempted to fit a *full model* with the formula *lmer(metric~ age + (1|corpus/child))*. Children in our combined corpus vary widely in age (0-78 months), and therefore some differences across children are actually due to their age. Therefore, we include age as a fixed effect to account for this variance. Similarly, it could be that children in our combined corpus differ because they come from different corpora, a proxy for differences due to the fact that they are learning different languages, growing up in different cultural settings, or even between-lab differences in terms of when parents were asked to record. We account for this common source of variance across children belonging to the same corpus by nesting child within corpus, and therefore declaring random intercepts for them.

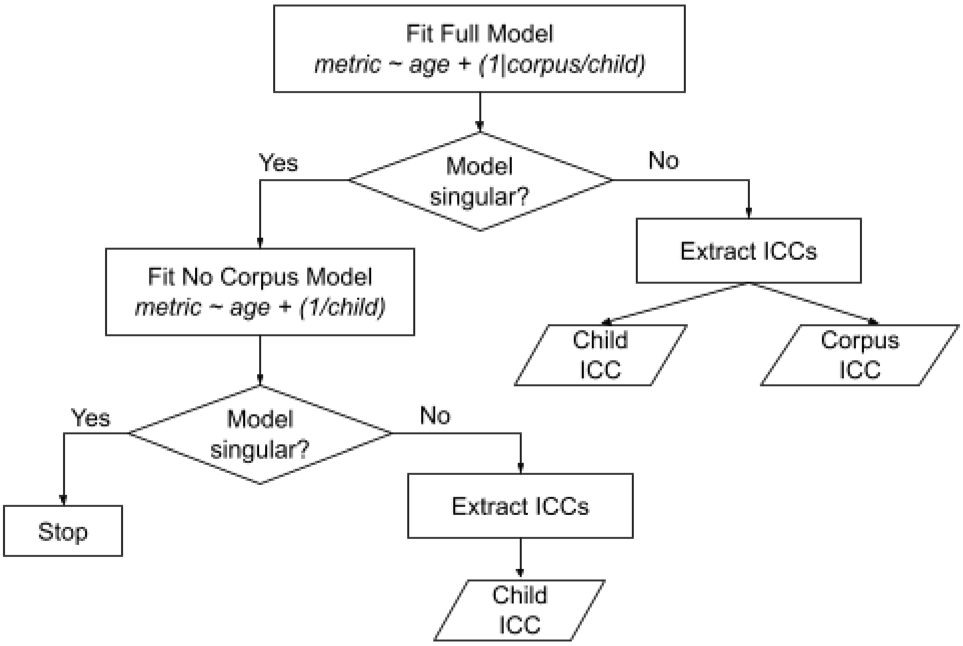

Figure 2. Model fitting process. Notice that when the "No Corpus Model" is singular, that means that variance cannot be attributed to individual variation across children once age has been controlled for. ICC cannot be extracted from singular models.

Note that our full model does not account for differences in development in group trajectories (i.e., differential age effects within corpus) or in individual trajectories (i.e., differential age effects within child). This is because when we attempted to fit more complex random structures (declaring age as a random slope within corpus and/or child), the majority of models were singular. We interpreted this as a sign that the model was overfitted, i.e., we cannot properly attribute variance to the predictor variables, either because a variable included does not explain any meaningful variance (e.g., we do not have enough individual infants' data points to fit a random slope for age by each individual child) and/or because it does not explain variance above and beyond another variable (i.e., the model cannot establish whether some variance is due to age differences present across all children versus within each group of children).[6]

Even this simpler random structure (child nested within corpus as random intercept) led to a singular model in a small number of cases. Given that this happened seldom (as reported in the Results section), we decided not to exclude the metric for which it happened, but rather simplified the random effects structure by fitting a random intercept by child using the following formula: *lmer(metric~ age + (1|child))*. From now on this model will be referred to as the *no corpus* model. This resolved the convergence problem, suggesting that in such cases corpus did not account for meaningful variance and/or did not do so above and beyond variance explained by individual children. Note that when *no corpus* model is

[6] In our analyses, the fixed effect of age should capture effects that are consistent across children as a function of the age the child has at the time of the recording. We did not find a way to capture potential effects of the amount of time that has elapsed across recordings or how such effects themselves might change with age (e.g. a 1 month age difference at 1 month is more extreme than a 1 month difference at 12 months). However, data are shared and may be analyzed by others to control for this variance.

singular, explained variance is due to age only, and not to between-child differences. Because our interest is in the proportion of variance that can be attributed to the child, and ICC cannot be extracted from singular models, such models cannot be incorporated in our analyses. This never happened for our main analyses, but it did for the exploratory analyses that we explain next.

In our main analyses, including all data means that repeated recordings within the same child could span a long time and many age ranges. For instance, in the Lucid dataset, children were recorded repeatedly between 10 and 31 months, so Child ICC captures relative stability across a very long time span, both in infancy and early childhood. It may be that metrics are reliable along shorter time scales of a few months, but not long time scales of over a year; and metrics are more reliable at some ages than others. We addressed both of these in exploratory analyses that subset data into 6-month age bins, as explained in more detail below. Model fitting proceeded similarly to Figure 1, although as we explain below, more models lead to singularity due to the limited number of data points. Finally, another exploratory analysis estimates reliability within individual corpora.

Results

Fuller output of analyses (including ANOVA tables for regressions) can be found in the online supplementary materials, available from https://gin.g-node.org/LAAC-LSCP/RELIVAL/src/master/CODE/SM.pdf.

Setting the stage: Correlations based on paired analyses of recordings done within two months

To set the stage with analyses likely to be intuitive to readers, we first analyze a subset of our data using correlations. Since correlations require paired observations, we subset our data to include two recordings that were less than two months apart, which allows us to compare our results with those of Gilkerson et al. (2017), who similarly resampled their longitudinal recordings. Our process is illustrated in Figure 3. Out of 191 children in 8 corpora, 148 children (belonging to 7 corpora) could be included in this analysis, as some children did not have recordings less than two months apart. In particular, no child from the Warlaumont corpus did.

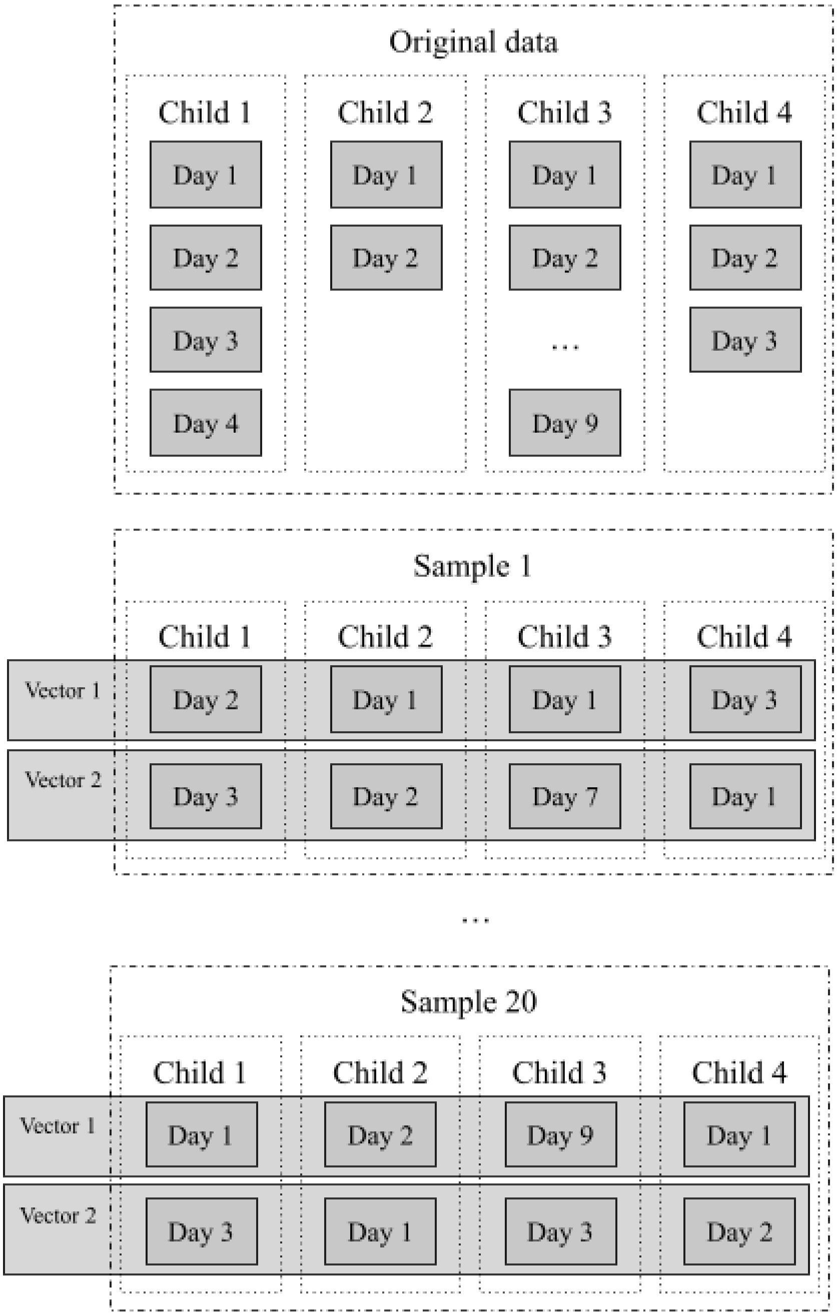

Figure 3. Illustration of the process whereby paired recordings are sampled. For each child who had several recordings less than two months apart, we randomly sampled one recording pair. We then computed the correlation between a given metric from the first sampled recording and the same metric from the second sampled recording across all children. For example, in Sample 1 shown here, the first vector is composed of e.g. the AWC for child 1 on day 2, child 2 on day 1, child 3 on day 1 and child 4 on day 3; the paired vector is the AWC for child 1 on day 3, child 2 on day 2, child 3 on day 7 and child 4 on day 1. Then the correlation coefficient corresponds to the level of association across the two sampled days, paired by child (i.e., child 1's days 2 and 3, child 2's days 1 and 2, etc.) To be able to describe potential variance emerging from particular recordings being randomly sampled and paired, we repeated this process 20 times, which led to more stable averages than 10 repetitions. If a child only has 2 recordings, then every sample will have the same two recordings for this child.

Table 3: Mean [range] of Pearson correlations across sampled pairs of recordings gathered less than two months apart (see Figure 3 for illustration of the process), along with reported mean r value by Gilkerson et al. (2017) for recordings done one or two months apart.

	ACLEW	LENA		
	Present study		Gilkerson et al. 2017	
	< 2 months apart		4 weeks apart	8 weeks apart
AWC per hour	.55 [.47,.63]	.52 [.40,.64]	.44 [.32,.54]	.40 [.28,.51]
CVC per hour	.80 [.76,.84]	.70 [.64,.76]	.69 [.61,.76]	.64 [.56,.72]
CTC per hour	.76 [.72,.80]	.69 [.61,.77]	.71 [.64,.77]	.66 [.58,.74]
Number of child vocalizations per hour	.69 [.63,.75]	.60 [.52,.68]	NA	NA

Table 3 shows the correlation values in our dataset for a handful of selected variables, comparable to the ones reported in Gilkerson et al. (2017). Focusing on our LENA column, and comparing it against the two columns for Gilkerson, we observe that the results for our LENA analysis tend to be numerically different from Gilkerson's, with correlations for AWC being higher in our data, whereas for the others our correlations are intermediate with respect to the 4- and 8-week analyses by Gilkerson. However, since the ranges overlap widely across our data and Gilkerson's, differences are not statistically significant.

Figure 4 shows the general distribution of correlation values, including all 75 metrics. To see whether correlations in this analysis differed by talker types and pipelines, we fit a linear model with the formula *lm(cor ~ type*pipeline)*, where type indicates whether the metric pertained to the key child, (female/male) adults, other children; and pipeline LENA or ACLEW. The model was significant overall (F(65) = 3.74, $p < .001$), with an adjusted R-squared of 25%. A Type 3 ANOVA on this model revealed a significant effect of type (F = 7.42, $p < .001$), with higher correlations for other children metrics (M = 0.66, SD = 0.1) and output metrics (M = 0.61, SD = 0.11) than female (M = 0.48, SD = 0.12), and male (M = 0.46, SD = 0.06) metrics. There was also a marginal effect of pipeline (F = 3.28, p = 0.07), with numerically higher correlations for ACLEW (M = 0.57, SD = 0.13) than for LENA metrics (M = 0.53, SD = 0.12). See SM G for fuller results.

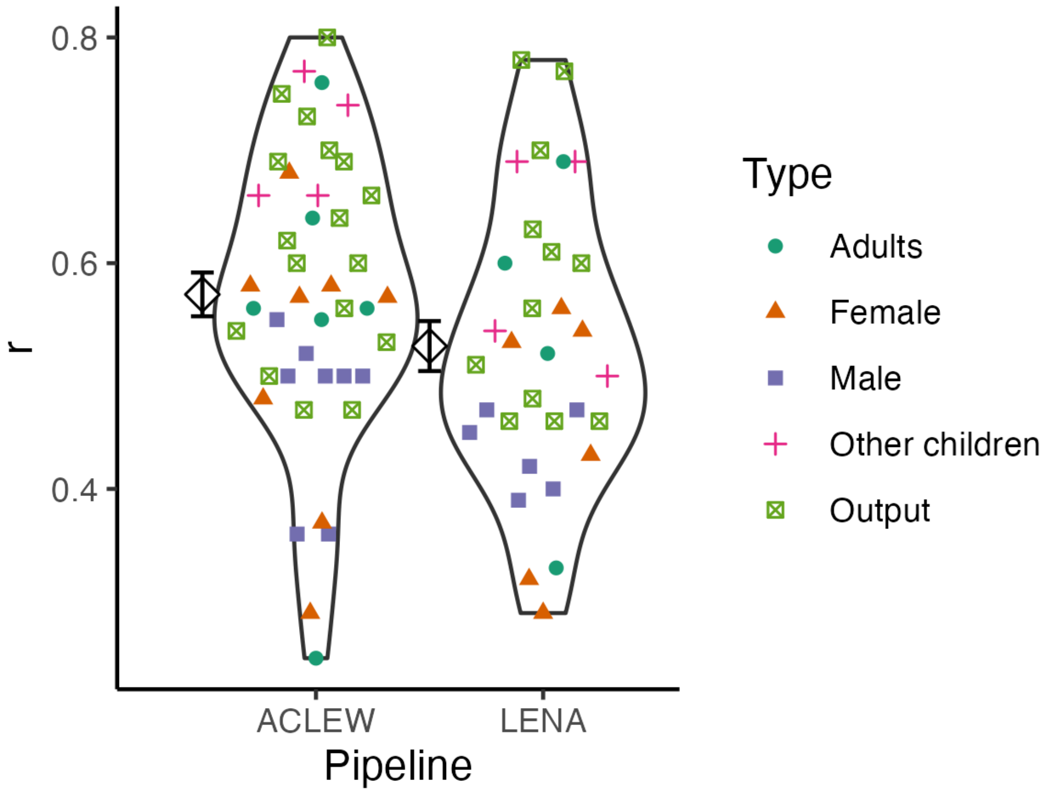

Figure 4. Violin plot reflecting the distribution of correlations across paired data points randomly sampled within each child that had recordings less than two months away from each other. Each point is a metric (see Table 1 for full list of metrics), with color indicating whether the metric pertains to vocalizations by the key child (output), female adults, male adults, adults collapsing across gender, or other children. Points have been jittered along the horizontal axis to facilitate inspection. The width of the violin indicates how many observations fall at a certain level. The white diamonds indicate the means for each violin plot.

This correlation-based analysis is suboptimal because we cannot benefit from all the extant data, since we must include exactly two points per child. In addition, the correlations do not control for variance attributable to differences across children due to age and corpora. When we do not explicitly control for this variance, all differences across corpora and across children (notably including their age) are captured through the correlational estimate. This could result in inflated coefficients, particularly for metrics that describe children's output (as shown in SM I). Therefore, the following analyses focus on Child ICC from mixed models.

Overall reliability

To begin with, we report on the Child ICC for metrics that may be most interesting to readers based on their popularity, shown in Table 4. In this particular subset of 5 metrics, it is not the case that LENA or ACLEW pipelines lead to systematically higher or lower Child ICCs, suggesting the choice of pipeline does not strongly affect the reliability of a study provided one of these 5 metrics is employed.

Table 4: Child ICC for the most commonly used metrics, namely LENA's AWC, CVC, CTC; and conceptually similar ones from both LENA and ACLEW.

	LENA	ACLEW
AWC per hour	.37	.33
CVC per hour	.43	.42
CTC per hour	.35	.46
Number of female vocalizations per hour	.39	.37
Number of child vocalizations per hour	.42	.45

Is it the case that other metrics may have even higher ICCs? Out of the 75 fitted models, 71 could be fit with the full model, yielding a measure of Corpus ICC. For the 4 for which the full model was singular, we fit the data with the No Corpus model, and none was singular then, allowing us to have Child ICC for all 75 metrics. The Child ICC of the more commonly used metrics shown in Table 4 hover around the average ICC of all 75 metrics; see the overall distributions of Child ICC illustrated in Figure 5. Indeed, the majority of metrics had Child ICCs between .3 and .5. Only seven metrics had Child ICCs higher or equal to .5. Surprisingly, the top 6 metrics in terms of Child ICC corresponded to the "other child" category, reported to have the worst accuracy according to previous analyses (Cristia et al., 2020). In an analysis fully reported in supplementary materials (SM M), we find some evidence that this may be due to the presence versus absence of siblings.[7] The next metric with the highest Child ICC corresponded to an output measure, namely the total vocalization duration per hour extracted from ACLEW annotations, with a Child ICC of .5. Among adult metrics, the average vocalization duration for female vocalizations for ACLEW and the ACLEW equivalent of CTC had the highest Child ICC (.45 and .46, respectively).

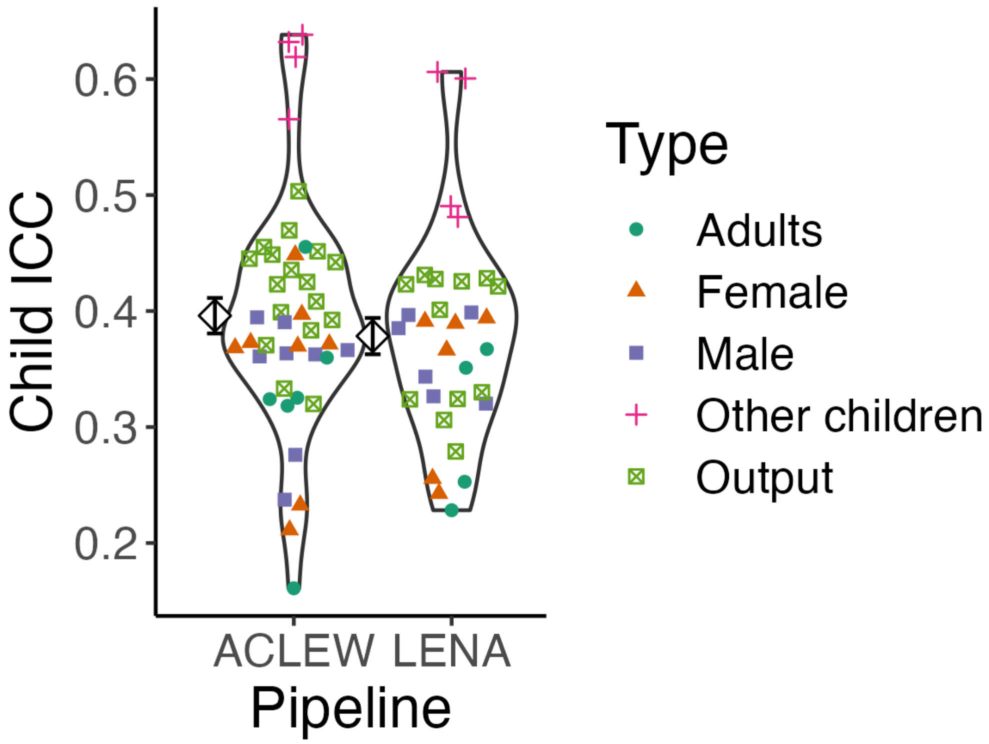

Figure 5. Violin plot reflecting the distribution of Child ICC.

[7] Along similar lines, in SM N, we document that a group of measures found to be relatively inaccurate when benchmarked against human data (particularly those splitting key child vocalizations into cry, canonical, and non-canonical) nonetheless results in average Child ICCs that hover around the average ICC of all 75 metrics.

Next, we explored how similar Child ICCs were across different talker types and pipelines. We fit a linear model with the formula *lm(icc_child_id ~ type * pipeline)*, where type indicates whether the metric pertained to the key child, (female/male) adults, other children; and pipeline is either LENA or ACLEW. The model was overall significant (F(65) = 12.11, p < .001). We found an adjusted R-squared of 57%, suggesting much of the variance across Child ICCs was explained by these factors. A Type 3 ANOVA on this model revealed type was a significant predictor (F(4) = 26, p<.001), and pipeline was marginal (F(1) = 2.8, p = 0.097); the interaction between type and pipeline was not significant. The main effect of type emerged because output metrics tended to have higher Child ICC (M = .4, SD = .06) than those associated to adults in general (M = .31, SD = .08), females (M = .34, SD = .07), and males (M = .35, SD = .05); whereas those associated with other children had even higher Child ICCs (M = .58, SD = .06). The trend for a main effect of pipeline arose because of slightly higher Child ICCs for the ACLEW metrics (M = .4, SD = .1) than for LENA metrics (M = .38, SD = .09). See SM P for fuller results.

Reliability across age groups

An open question is whether reliability can be generalized from one age range to another. To assess stability of Child ICC as a function of child age, we re-did analyses splitting observations as a function of the child age in 6 month bins. That is, we took data from all corpora in a given age bin (say 0 to 6 months), then fit a model with the flowchart in Figure 1 to extract Child ICC. We only did so when there were at least 30 observations for a given age bin, which means not all age bins in the data are represented in this analysis.

Out of 450 fitted models (75 metrics times 6 age bins), 6 were singular when including a random intercept per child, and therefore they could not be included in these analyses at all. In addition, 146 were singular when including a random intercept per corpus. The remaining 298 could be analyzed with the full model.

Figure 6 shows the distribution of Child ICC in each age bin, as well as the number of included children and corpora. The most salient effect may be the fact that Child ICC tends to be considerably less variable across metrics for the 6-12 and 12-18 month bins. We also had more children and different corpora represented in those bins. In these subsets, most metrics have Child ICCs between .3 and .5, whereas in other ages we see more variation, with a relatively greater proportion of metrics with Child ICC lower than .3 than in both younger bins (0-6 months, or above 24 months). Another observation is that the very high values of Child ICC for metrics pertaining to other children is most obvious for the 0-6 and 6-12 month bins, where these metrics have Child ICCs above .6. In contrast, above 18 months of age, other children metrics have relatively high Child ICC (above .5), but so do metrics pertaining to the key child.

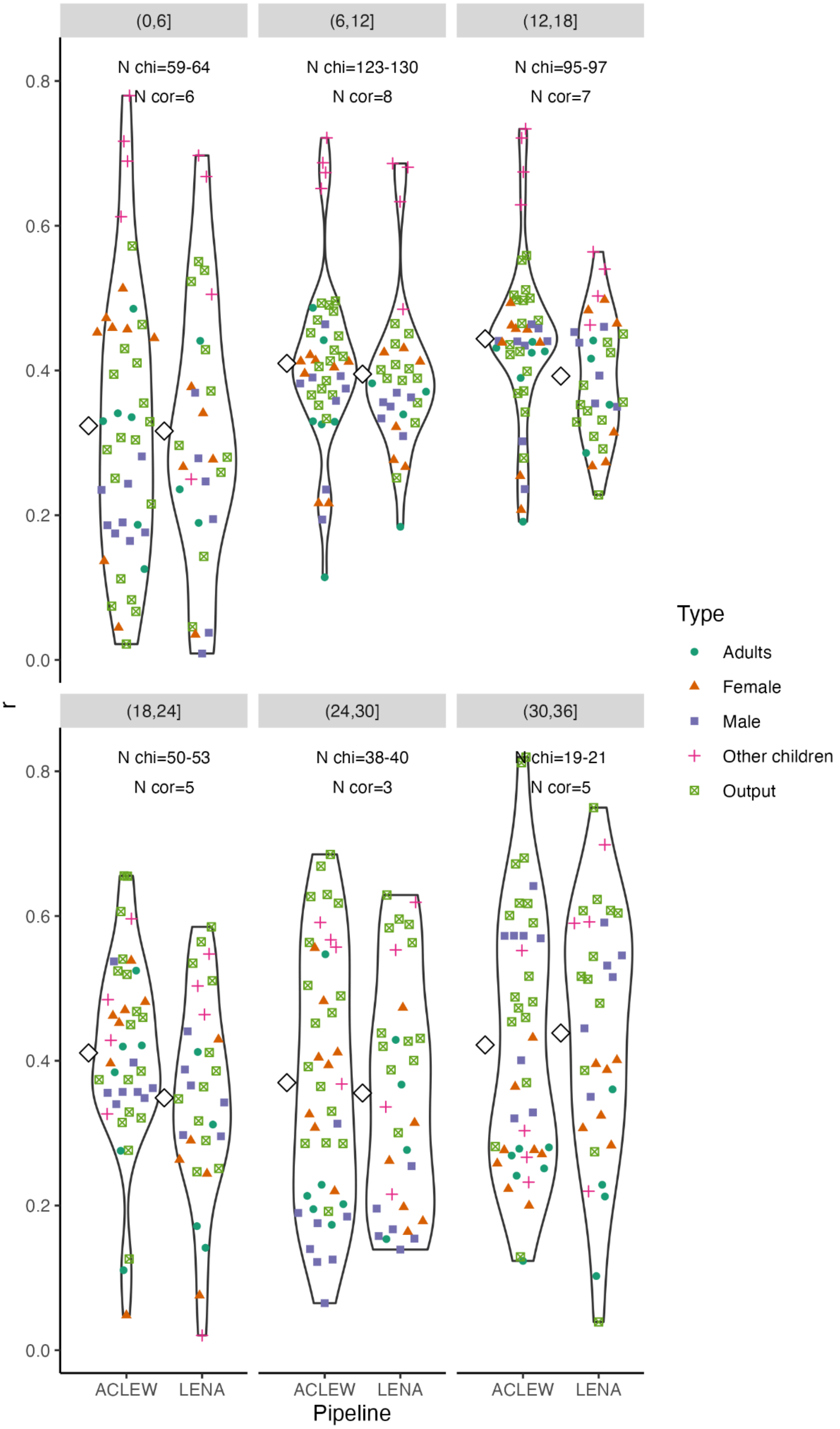
(0,6]
(6,12]
(12,18]
N chi=59-64
N cor=6
N chi=123-130
N cor=8
N chi=95-97
N cor=7
(18,24]
(24,30]
(30,36]
N chi=50-53
N cor=5
N chi=38-40
N cor=3
N chi=19-21
N cor=5
0.0
0.2
0.4
0.6
0.8
r
ACLEW
LENA
Pipeline
Type
Adults
Female
Male
Other children
Output

Figure 6. Child ICC by metric type and pipeline, when considering each age bin separately. N chi indicates the number of children that could be included (which varies across metrics). N cor indicates the number of corpora that could be included. The circled symbols to the right of each violin indicate the mean Child ICC for that metric as a function of type.

To interrogate these results statistically, and assess whether Child ICCs tended to be higher or lower in certain age bins, we fit a linear model with the formula *lm(icc ~ type * pipeline * age_bin)*. The model was overall significant (F(372) = 5.83, p < .001). We found an adjusted R-squared of 40%, suggesting this model explained more than a third of the variance in Child ICC. A Type 3 ANOVA on this model revealed that the interactions between type and pipeline, pipeline and age, and the three-way interaction (type, pipeline, age) were not significant. However, both the type*age bin interaction (F(20) = 6.7, p < .001) and the three main effects were significant (type: F(4) = 30.4, p < .001; age: F(5) = 8.6, p < .001; pipeline: F(1) = 8.6, p = .01). The significant interaction between type and age bin captures the wide variation in Child ICC as a function of these two factors: for example, notice that other child metrics are much higher (above .5) for 0 to 18 months and average at 18-24 months. See SM S for more information.

Finally, we assessed whether one can generalize from Child ICCs found in one age bin to other age bins. To this end, we calculated a correlation matrix for the Child ICC in paired observations across every pair of age bins (e.g., the Child ICC obtained within 0-6 months bin against the Child ICCs obtained for the same set of metrics and pipelines for the 6-12 months bin). This allows us to capture the idea that the relative scores of Child ICC (i.e., how high Child ICC is for one metric compared to another metric) are stable across age bins. Results are shown in Figure 7. This analysis suggests that generalization across age bins is possible because correlation coefficients are generally positive and moderate in size, with most coefficients between .22 and .5. Another obvious result is that correlations are lowest when they involve the 30-36 month bin, which is the one with the smallest number of included children, and thus may not be representative. One caveat is in order: If many children recur across bins, this may lead to cross-bin correlations that are higher than they would be if participants are not shared across bins. We have not devised a way to address this issue, but share all data in the hope that others may be able to do so.

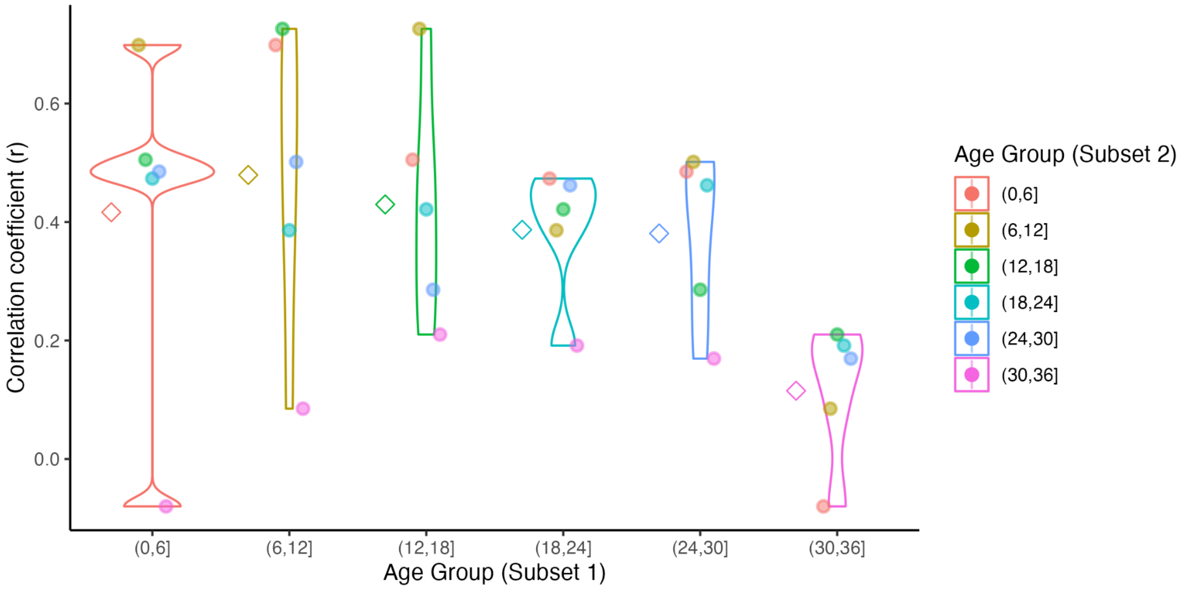

Figure 7. Correlations in Child ICC across age bins. Each point indicates the correlation in Child ICC for the age bin named in the x-axis with every other age bin.

Reliability within corpus

Readers may wonder whether they should inspect Child ICC in their own data set as a way to select the most reliable metric for a given construct. This may not be necessary if the most reliable measures in one corpus are also the most reliable ones in other corpora. From the perspective of generalizability of findings, it is preferable to use the same metrics across corpora - therefore it is of interest to examine the extent to which Child ICC varies across corpora. We investigated this by fitting models within corpora, but we note here that 6 out of the 8 corpora have fewer than 30 children, which means results may be unstable due to small sample sizes.

Out of 600 fitted models (75 metrics times 8 corpora), 26 were singular when including a random intercept per child, and therefore they could not be included in these analyses at all. (Including a random intercept per corpus is not relevant here, since only data from one corpus is included in each model fit.) Figure 8 shows the distribution of the resulting Child ICC values. There are similarities across corpora, with most metrics in the .3-.5 region and largely overlapping distributions. However, there are differences across corpora as a function of pipeline: Whereas in most other corpora the distributions for LENA and ACLEW are overlapping, much of the mass in the Child ICC distribution for LENA is below .5 in Quechua specifically. Visual inspection also reveals three-way interactions, since the ranking of Child ICC by type can differ markedly across corpora. One example is that other child metrics have the highest Child ICCs in several corpora (i.e., bergelson, lucid) but low ICCs for others (in particular, for winnipeg ACLEW).

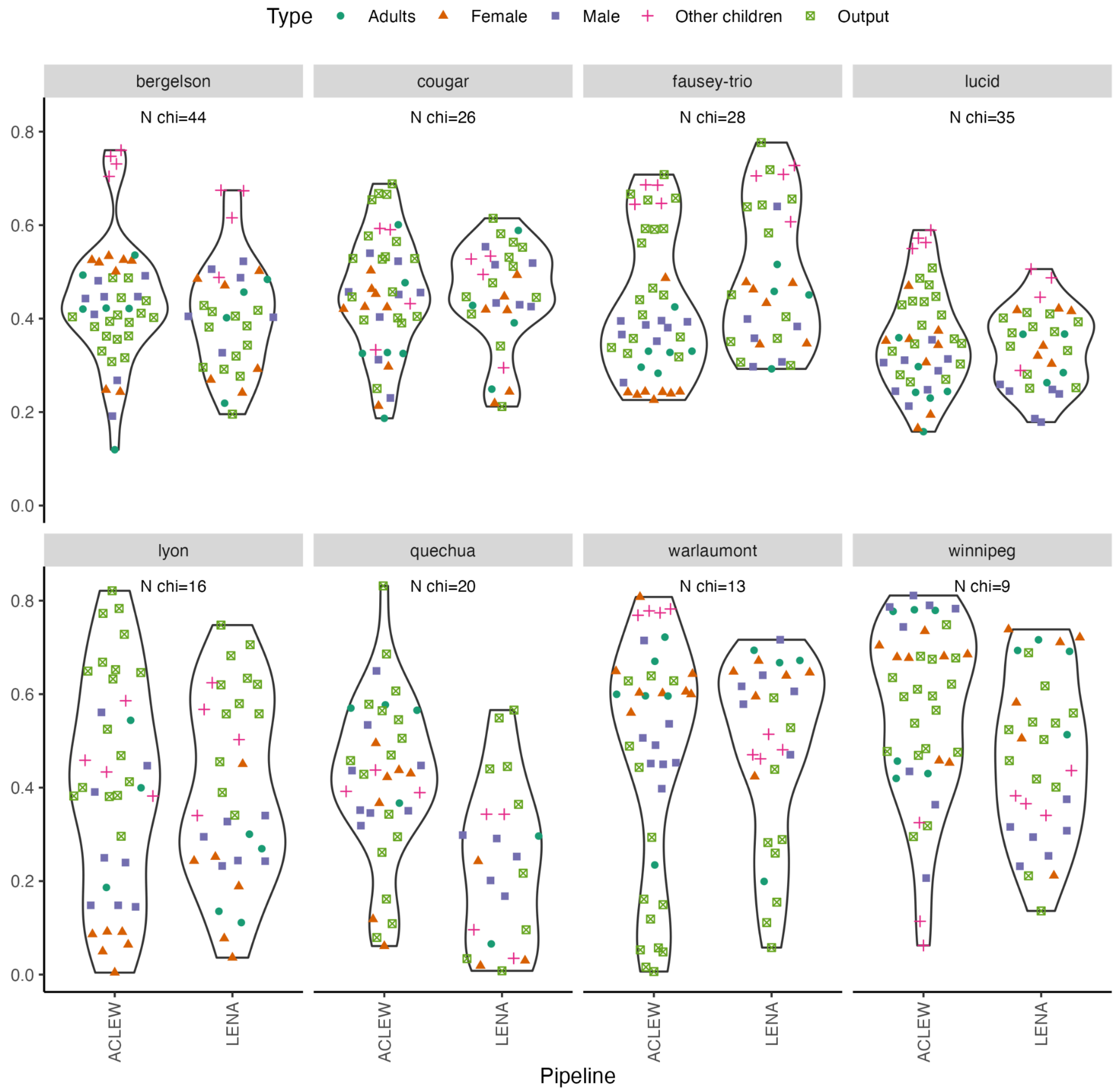

Figure 8. Child ICC by metric type and pipeline, when considering each corpus separately.

The fact that we cannot infer reliability from one corpus based on another one was confirmed statistically: We checked whether Child ICC differed by talker types and pipelines across corpora by fitting a linear model with the formula *lm(icc ~ type * pipeline * corpus)*, where type indicates whether the measure pertained to the key child, (female/male) adults, other children; pipeline LENA or ACLEW; and corpus the corpus ID. The model was overall significant (F(494) = 7.42, p < .001). We found an adjusted R-squared of 47%, suggesting this model explained nearly half of the variance in Child ICC. A Type 3 ANOVA on this model revealed several significant effects and interactions, including a three-way interaction of type, pipeline, and corpus (F(28) = 2, p<.001); a two-way interaction of type and corpus (F(7) = 4.9, p<.001); and a main effect of corpus (F(7) = 16.5, p<.001).

As with age, we calculated a correlation matrix in terms of Child ICC, this time across corpora. Results are shown in Figure 9, where we see that, in fact, the ranking across metrics as a function of Child ICCs found within one corpus are only weakly correlated, and can even be negatively correlated, to the same in another corpus, with the clearest example of this being Winnipeg. We believe this is not a feature of the Winnipeg corpus per se, but rather it is a possible outcome from the fact that the Winnipeg corpus is quite small (it contains only 9 children), so that reliability is not accurately estimated: It could be underestimated due to noise, or overestimated (for instance, if there is structured variance in the corpus beyond child age; e.g., if some of the children were consistently recorded in daycares and others at home). Overall these findings suggest that the relative reliability of different metrics is not stable across corpora, suggesting caution before assuming that the level of reliability of a metric can generalizes across corpora. Although this is based on the relative ranking of reliability, this follows from variation in the absolute levels of reliability. For instance, Child ICC for AWC varied between .27 and .72 across corpora.

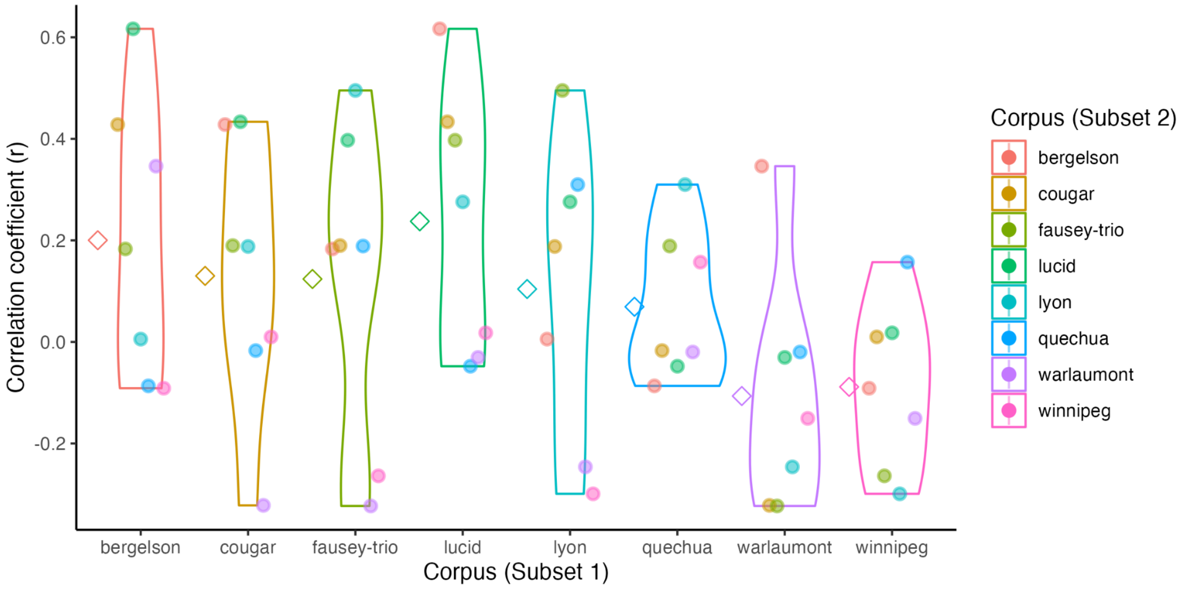

Figure 9. Correlations in Child ICC across corpora. Each point indicates the correlation in Child ICC for the corpus named in the x-axis with every other corpus as measured across all the 75 metrics.

Discussion

This study examined the reliability of 75 different metrics extracted from long-form audio recordings using two pipelines (LENA and ACLEW). A first analysis using paired recordings (collected less than two months apart from each other) showed that children's relative scores within a given metric yielded correlations between .5 and .8, and those for AWC, CVC, and CTC were in line with previous reports (d'Apice & Stumm, 2019; Gilkerson et al., 2017). Using our mixed model analyses, which allowed us to include all of the data, we find that reliability is lower than what those correlation estimates would have led us to expect, with Child ICCs between .2 and .5. We believe this is because Child ICCs from mixed models control for differences in child age (and corpus, when this accounts for some variance), as simulations suggest that Child ICC values can be read similarly to correlation values (see footnote 1 and SM A). The reliability of .2-.5 found for our metrics is lower than values expected from three meta-analyses, one on reliability of the Ages and Stages parental questionnaire finding rs > .6 (Denman et al., 2017), another on reliability of standardized language tests on older children's behavior with average r = .67 (Denman et al., 2017), and the third on behavioral measures collected from wearables among adults tested in laboratory conditions leading to ICCs above .6 (Kobsar et al., 2020). We may observe lower reliability in the present study because we are looking at infants, whereas those meta-analyses pertain to children and adults respectively. However, it is also important to bear in mind that both of those meta-analyses warned against the low quality of the data included, which may indicate that they are overestimating reliability. Moreover, those meta-analyses are only valuable as approximate benchmarks for what we might expect to find, rather than directly comparable analyses, given the lack of prior studies of this type. Our reliabilities are also lower than those observed for wide-spread parental report instruments like the MB-CDI (SM B), but as argued in the Introduction, the latter reflect stability not only in the infant but also in the parent. Moreover, the MB-CDI is more limited in scope in what it captures of the child's experience and behaviors relative to longform audio metrics. In contrast, the levels of reliability we observe here for longform audio metrics are much greater than those observed in previous infant research focusing on laboratory tasks, which in a meta-analysis shockingly yield a weighted mean correlation of zero (Cristia et al., 2016).

In addition to inspecting overall levels of reliability, our analyses attempted to provide useful information regarding relative differences in reliability as a function of the pipeline and metric. Regarding pipeline, the choice between LENA and ACLEW may be of notable interest to readers, and it was not obvious which of these would lead to greater reliability. On the one hand, since most included data was collected from children learning American English (5 out of 8 corpora), LENA could have been at an advantage because LENA algorithms were trained exclusively on American English, whereas the ACLEW algorithms were trained on a diverse set of languages, with English constituting only a subset of the training set. On the other hand, when the segmentation into talker types of LENA and ACLEW was benchmarked against human annotations, ACLEW had higher accuracy than LENA (Lavechin et al., 2020). One could predict higher accuracy should allow higher reliability. Although some of our analyses suggest a statistically significant advantage for ACLEW, the difference is very small, and some specific metrics that exist in both pipelines have numerically higher Child ICC for LENA than ACLEW. Moreover, for the most commonly employed metrics, the difference is very slight. One conclusion from this is that the free, open-source alternative algorithm developed in the ACLEW project hence appears as a promising alternative to the closed-source LENA algorithm. However, as discussed in previous tutorials for long-form recordings, LENA is both easier to use and better known (Pisani et al., 2021). Therefore, our conclusion is that researchers may not be greatly disadvantaged by using either of the pipelines, depending on the needs and available resources in their lab.

There is another aspect of the pipeline comparison that invites pause. As a reminder, accuracy indicates the extent to which the algorithm agrees with human annotations for the exact same sections of the audio, whereas reliability indicates relative stability in children's rankings for their input or output metrics. We were surprised by the fact that accuracy differences did not translate into reliability differences. This observation also emerged when we inspected possible differences between metrics. Most saliently, previous research strongly suggests that LENA's AWC and CVC are a great deal more accurate than CTC, with meta-analytic mean r's of .79, .77, and .36 respectively (Cristia et al., 2020). Yet their reliabilities were very similar, at Child ICC = .37, .35, and .32 respectively.

Another case in which differences in accuracy did not translate into reliability differences comes from an inspection of the metrics having the highest Child ICC, which, surprisingly, are those associated with other children. In Lavechin et al., (2020), even the better-performing ACLEW algorithm shows stark accuracy differences between other and key child, with the other child category having an F-score of only 26%, whereas the key child has an F-score of 77%. This case highlights the importance of understanding the many factors that contribute to reliability: Test-retest reliability depends not only (and perhaps not primarily) on accuracy, but also (and probably crucially) on the stability of underlying conditions across recordings. In an analysis described in supplementary materials (SM M), we find evidence that the relatively higher Child ICC for other children's metrics may stem from the fact that children vary in whether they have siblings, because Child ICC was reduced when this factor was controlled for. Another observation that is consistent with this interpretation is that the Child ICCs associated with other children were lower for Quechua, the one rural dataset we included. According to the curator of the Quechua dataset, even children with no siblings in the Quechua corpus are exposed to diverse speakers, including child speakers, which may "wash away" differences in exposure to other children's speech. Moreover, their days may vary more (e.g., days when the child goes to the market with the mother versus others staying near the home), and thus in individual variation measured by other children's metrics. One aspect that we have not looked at, and neither has previous research, is the extent to which reliability (and Child ICC) vary as a function of patterns of algorithmic confusion: Perhaps the key child and other child are more frequently confused by the automated system when there are more children around.

Additional analyses checked whether there were systematic differences in Child ICC as a function of child age and corpora. We observed more consistent Child ICC levels for the two age bins in which we had most data, 6-12 and 12-18 months. Further study may be necessary to determine whether age or sample size are more at play in this finding. We had expected adult behavior to be most reliable and stable, but relatively higher reliabilities were observed for metrics pertaining to children rather than adults, particularly among older cohorts. This may indicate that there is more stability in children's spontaneous behavior at older ages, whereas perhaps the child's quantity and quality of vocalization may be more affected by random variation in, for instance, their mood.

Regarding corpora, a recent analysis suggested that, at least for two metrics (adult and child vocalization counts), accuracy did not systematically vary across English and non-English, urban and rural settings, but it did vary markedly across corpora in ways that have yet to be explained (Bergelson et al., 2023, see supplementary material). We observed cross-corpus differences in the overall spread of Child ICC, and detected statistically significant differences in Child ICC as a function of pipeline and metric type across corpora. Differences across corpora could indicate that the present results may not readily generalize to the corpora collected by readers. However, we think this conclusion may be premature, as some of the divergence seems to be due to the fact that corpora varied in terms of the number of participants, with none of the corpora being particularly large. We thus place more credence in the overall reliability analyses, which benefited from a much larger dataset.

Some readers may be interested in selecting one highly reliable metric for children's production and another for adult input. We propose that metric choice takes into account reliability, but also the concept thought to be captured by the metric. For children's production, the highest Child ICCs were obtained not with LENA's CVC, but with an ACLEW measure that additionally takes into account children's vocalization duration. Conceptually, this seems like a reasonable metric to use, as it represents cumulative duration of vocalizations produced by the child wearing the device. Its reliability may be higher than that of other output metrics for a range of reasons that should be explored in future research, including empirical and methodological ones. As for empirical reasons, perhaps children's volubility (speech quantity) is more stable than qualitative aspects of their spontaneous production. Regarding methodological reasons, it may be the case that a measure that cumulates duration reduces random methodological noise more effectively than one based exclusively on counts. Among adult metrics, the highest ranking again came from the ACLEW pipeline, and it was the average female vocalization duration. Average vocalization duration seems closer to mean length of utterance than cumulative input experienced. We similarly hope that future work investigates why this metric is relatively more stable.

The present study has a salient limitation: Our data draws mainly from urban (96% of recordings, 90% of the children, 88% of the corpora), English-speaking settings (92% of recordings, 81% of the children, 75% of the corpora), largely from North America (73% of recordings, 63% of the children, 62% of the corpora). We only included children who had normative development (according to parental report), and the majority of the datasets are also biased towards relatively high SES families, which may limit the extent of individual variation present in our dataset. For readers who are interested in group differences across children varying in normativity and/or other individual or familial factors, this could mean that our reliability is underestimated (since we would expect a wider range of child speech abilities in, e.g. populations with and without language-relevant clinical diagnoses). Moreover, we only considered data collected with LENA recorders since we wanted to have head-to-head comparisons of LENA and ACLEW algorithm outputs, and LENA software can only be applied to recordings gathered with LENA hardware which requires substantial resources, usually including regular access to an internet connection. As the field starts employing other recording devices, it will be crucial to consider variation related to hardware as well (as argued in SM Y).

Future work may also consider consistency according to day of the week, also taking into account that in some corpora weekdays may differ from weekends, in others only the Sunday may be different, or variability could be due to a host of additional conditions, including holidays, climate, season, and more. Additionally, while we included a large set of metrics linked to both children's output and input, we did not explore whether metrics that are composite across several of these (e.g., through principal component or factor analysis) may display higher reliability. For instance, it is possible that a measure of "vocalization development" that combines vocalization count, duration, and the proportion of vocalizations that are canonical may display higher reliability than each of these. More broadly, we believe that a fruitful avenue of work would be to employ factor analysis and other such techniques to study convergent and divergent validity of the metrics studied here, and others proposed in the future. Finally, we focused on reliability from metrics based on minimally 4h of audio data, but it may be that even one full day does not suffice to gain a representative snapshot of children's language environment. As hardware develops, researchers may be able to more easily acquire multiple consecutive days of recordings, thus being able to average out atypical days. We expect reliability will increase with increases in data representativeness.

Conclusions

This study is the first large-scale analysis of the test-retest reliability of metrics derived from automated annotations of long-form recordings. This collaborative effort included data from 8 corpora collected in 4 different countries, including a total of 1,254 day-long recordings from 191 children. Results suggested that metrics varied markedly in reliability. Reliability estimates were in the low to moderate range, corresponding to correlations of .2-.5; they were furthermore decoupled from metrics' accuracy (established in comparison to manual annotation). One of the highest reliability metrics was cumulative duration of key child vocalizations, and the most stable reliability levels were observed for children between 6 to 18 months. There were some differences across audio analysis pipelines, with certain statistical analyses favoring ACLEW-based metrics, but these were numerically small. The present work provides a solid basis for future research considering variability in underlying behavior as well as the validity of automated metrics, which would ideally be carried out on geographically and normatively diverse samples.

Open Practices and Data availability statement

Anonymized (tabular) data and all relevant code required to reproduce this manuscript have been deposited in GIN (https://gin.g-node.org/laac-lscp/relival). None of the experiments was preregistered.

References

Al Futaisi, N., Zhang, Z., Cristia, A., Warlaumont, A., & Schuller, B. (2019). VCMNet: Weakly Supervised Learning for Automatic Infant Vocalisation Maturity Analysis. *2019 International Conference on Multimodal Interaction*, 205–209. https://doi.org/10.1145/3340555.3353751

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), Article 1. https://doi.org/10.18637/jss.v067.i01

Bergelson, E. (2017). *Bergelson Seedlings HomeBank Corpus*. doi:10.21415/T5PK6D

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. https://doi.org/10.1111/desc.12715

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., Alphen, P. van, & Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENA^{TM}) in European French. *Behavior Research Methods*, *48*(3), 1109–1124. https://doi.org/10/f83h7j

Casillas, M., & Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology*, *5*(1), 24. https://doi.org/10/gnzf8w

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis System Segmentation and Metrics: A Systematic Review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–Retest Reliability in Infant Speech Perception Tasks. *Infancy*, *21*(5), 648–667. https://doi.org/10.1111/infa.12127

Cychosz, M. (2022). Language exposure predicts children's phonetic patterning: Evidence from language shift. *Language*. https://doi.org/10/grgntc

Cychosz, M., Edwards, J., Munson, B., Romeo, R. R., Kosie, J., & Newman, R. (2023). *The everyday speech environments of preschoolers with and without cochlear implants*. https://doi.org/10.31234/osf.io/kvzt4

d'Apice, K., & Stumm, S. von. (2019). *Does Age Moderate the Influence of Early Life Language Experiences? A Naturalistic Home Observation Study*. PsyArXiv. https://doi.org/10.31234/osf.io/jr4by

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y.-W., & Cordier, R. (2017). Psychometric Properties of Language Assessments for Children Aged 4–12 Years: A Systematic Review. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01515

Drude, S., Broeder, D., Trilsbeek, P., & Wittenburg, P. (2012). The Language Archive – a new hub for language resources. *LREC*, 3264–3267.

Fausey, C. M., & Mendoza, J. K. (2018). *FauseyTrio HomeBank Corpus* [dataset]. https://doi.org/10.21415/T5JM4R

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), 1–173; discussion 174-185.

Fibla Reixachs, L. (2021). *Relating language input to language processes early in development: Using the early language processing task in UK and India* [Doctoral, University of East Anglia. School of Psychology,]. https://ueaeprints.uea.ac.uk/id/eprint/83017/

Ganek, H., & Eriks-Brophy, A. (2018). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. *Journal of Communication Disorders*, *72*, 77–85. https://doi.org/10/gmtkpd

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. https://doi.org/10/gfzjg3

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a Psychological Theory: Integrating Construct-Validation and Computational-Modeling Methods to Advance Theorizing. *Perspectives on Psychological Science*, *16*(4), 803–815. https://doi.org/10/ghtczb

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.

https://doi.org/10/fgpwcm

Kobsar, D., Charlton, J. M., Tse, C. T. F., Esculier, J.-F., Graffos, A., Krowchuk, N. M., Thatcher, D., & Hunt, M. A. (2020). Validity and reliability of wearable inertial sensors in healthy adult walking: A systematic review and meta-analysis. *Journal of NeuroEngineering and Rehabilitation*, *17*(1), 62. https://doi.org/10.1186/s12984-020-00685-3

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *Interspeech*. http://arxiv.org/abs/2005.12656

Levin-Asher, B., Segal, O., & Kishon-Rabin, L. (2023). The validity of LENA technology for assessing the linguistic environment and interactions of infants learning Hebrew and Arabic. *Behavior Research Methods*, *55*(3), 1480–1495. https://doi.org/10.3758/s13428-022-01874-9

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). *performance: An R Package for Assessment, Comparison and Testing of Statistical Models*. *6*, 3139. https://doi.org/10.21105/joss.03139

McDivitt, K., & Soderstrom, M. (2016). *McDivitt HomeBank Corpus*. 10.21415/T5KK6G

Pisani, S., Gautheron, L., & Cristia, A. (2021). *Long-form recordings: From a to z*. https://doi.org/10.5281/zenodo.6685828

R Core Team. (2023). *R: A language and environment for statistical computing* (4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2021). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, *53*(2), 818–835. https://doi.org/10/ghb2g8

Rowland, C. F., Bidgood, A., Durrant, S., Peter, M., & Pine, J. M. (2018). *The Language 0–5 Project Corpus* [dataset]. https://doi.org/10.17605/OSF.IO/KAU5F

Schuller, B., Räsänen, O., Metze, F., Dupoux, E., & Cristia, A. (2024). *ACLEW tools report*. https://osf.io/nwc56/

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing a Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience. *Collabra: Psychology*, *7*(1), 23445. https://doi.org/10/gnzf8x

VanDam, M. (2018). *HomeBank English Cougar Corpus* [dataset]. https://doi.org/10.21415/T5WT25

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & Macwhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*, *37*(2). https://doi.org/10.1055/s-0036-1580745

Velikonja, T., Edbrooke-Childs, J., Calderon, A., Sleed, M., Brown, A., & Deighton, J. (2017). The psychometric properties of the Ages & Stages Questionnaires for ages 2-2.5: A systematic review. *Child: Care, Health and Development*, *43*(1), 1–17. https://doi.org/10.1111/cch.12397

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA^{TM} automated measures for child language development. *Developmental Review : DR*, *57*, 100921. https://doi.org/10/gmtkpq

Warlaumont, A. S., Pretzer, G. M., Mendoza, S., & Walle, E. A. (2016). *Warlaumont HomeBank Corpus*. doi:10.21415/T54S3C