Running head: CORPUS OF NATURALISTIC VOCALIZATIONS

Speech Maturity Dataset: A cross-cultural corpus of naturalistic child and adult vocalizations

Kasia Hitczenko[1], Loann Peurey[2], William N. Havard[3], Kai Jia Tey[2], Amanda Seidl[4], Chiara Semenzin[5], Camila Scaff[2,6], Marvin Lavechin[7], Bridgette Kelleher[8], Lisa Hamrick[8,9], Lucas Gautheron[10], Margaret Cychosz[11], Marisa Casillas[12], & Alejandrina Cristia[2]

[1] Department of Computer Science, The George Washington University, Washington, DC, USA

[2] Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Études Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

[3] Laboratoire Ligérien de Linguistique, University of Orléans, Orléans, France

[4] Department of Communication Sciences and Disorders, University of Delaware, Newark, DE, USA

[5] Institut de Biologie de l'ENS (IBENS), Département de biologie, Ecole normale supérieure, Université PSL, 75005 Paris, France

[6] Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland

[7] GIPSA-lab, Université Grenoble Alpes, Grenoble, France

[8] Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

[9] Department of Psychology, University of South Carolina, Columbia, SC, USA

[10] Interdisciplinary Centre for Science and Technology Studies, Wuppertal, Germany

[11] Department of Linguistics, University of California, Los Angeles, Los Angeles, CA, USA

[12] Department of Comparative Human Development, University of Chicago, Chicago, IL, USA

Author note

This work was funded by 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). This publication uses data generated via the Zooniverse.org platform, the development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. The funders had no impact on this study.

The authors made the following contributions. Kasia Hitczenko: Conceptualization, Software, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision; Loann Peurey: Conceptualization, Software, Data Curation, Writing - Review & Editing, Project Admin; William N. Havard: Conceptualization, Software, Data Curation, Writing - Review & Editing, Project Admin, Visualization; Kai Jia Tey: Software, Validation, Writing - Review & Editing; Amanda Seidl: Methodology, Validation, Data Curation, Writing - Review & Editing; Chiara Semenzin: Conceptualization, Validation, Software, Data Curation, Writing - Review & Editing; Camila Scaff: Data Curation, Writing - Review & Editing; Marvin Lavechin: Software, Data Curation, Writing - Review & Editing; Bridgette Kelleher: Validation, Data Curation, Writing - Review & Editing; Lisa Hamrick: Validation, Data Curation, Writing - Review & Editing; Lucas Gautheron: Methodology, Software, Data Curation, Visualization, Writing - Review & Editing; Margaret Cychosz: Methodology, Validation, Data Curation, Writing - Review & Editing; Marisa Casillas: Methodology, Data Curation, Writing - Review & Editing; Alejandrina Cristia: Conceptualization, Methodology, Validation, Data Curation, Writing - Review & Editing, Supervision, Funding. Authors are reverse-alphabetical, except for the first

three authors and the last author. We thank Gladys Baudet, Elika Bergelson, Heidi Colleran, Pauline Grosjean, Sarah Walker, Anne Warlaumont, and Lisa Yankowitz for contributing data and for their help at various stages of creating this dataset.

Correspondence concerning this article should be addressed to Kasia Hitczenko, 800 22nd St NW, Washington, DC 20052. E-mail: kasia.hitczenko@gwu.edu

CORPUS OF NATURALISTIC VOCALIZATIONS

Abstract

Over the first years of life, children's vocalizations become increasingly adult-like, and lay the foundation for later phonetic and phonological development. Yet research in this area has been limited to a narrow set of languages and communities, mainly Indo-European languages from Western(ized) speaker communities, and focused on a narrow age range (0-24mo). We present a new publicly-available dataset, the Speech Maturity Dataset, consisting of over 1 million labels, provided by over 20,000 unique annotators, of ~250,000 clips from long-form recordings of over 400 children (aged 1mo-6yrs) from 10 communities (ranging from rich industrialized societies to farmer-forager communities) speaking one or more of 41 languages. The clips include both (i) key child vocalizations labeled with their vocalization type (canonical vocalization, containing an adjacent consonant and vowel; non-canonical vocalization, with no adjacent consonant/vowel; laughing; or crying) as well as (ii) vocalizations from other speakers in the environment which are additionally labeled with speaker information (baby, child, female/male adolescent, female/male adult). This dataset, which includes child-level metadata (sex, age, monolingual status, diagnoses, etc.), can be used to study vocal development in an unprecedented way across a wide variety of communities. It can also be used to train vocalization-type classifiers in an effort to make software dedicated to the study of child language acquisition free, open-source, and reproducible. In sum, our dataset represents an ongoing and collaborative effort between field researchers, psycholinguists, and citizen scientists, and promises to allow scientists to significantly expand their knowledge of early vocal development.

Keywords: vocalizations; vocal development; language development; canonical vocalizations; citizen science; long-form recordings

Speech Maturity Dataset: A cross-cultural corpus of naturalistic child and adult vocalizations

1 Introduction

Over the first years of life, children's spontaneous vocal productions become increasingly adult-like (Oller, 2000). One pattern that has been documented is an age-related decrease in the prevalence of crying (Bergelson et al., 2023); another, is the increasing prevalence of fast consonant-vowel or vowel-consonant transitions, which has been captured through metrics like canonical babbling onset (Morgan & Wren, 2018) and canonical proportion (Cychosz et al., 2021). Past studies suggest continuity between the phonological inventory in babbling and first words (e.g., Majorano, Vihman, & DePaolis, 2014). Moreover, early vocal productions have been identified as a promising marker of speech and language disorders, and are a target for clinical interventions (Yankowitz, Schultz, & Parish-Morris, 2019). However, despite being extensive, past research in this area has been limited in a number of ways. First, children's vocal productions have primarily been studied using short samples taken in lab environments. This methodology skews participant samples demographically, with research over-representing North American, white, wealthy participants (Singh, Cristia, Karasik, Rajendra, & Oakes, 2023). Moreover, shorter in-lab recordings may also misrepresent infants' production; for instance, one study found infants vocalized more at home than in the lab (Lewedag, Oller, & Lynch, 1994). Finally, past research has used different measures for younger versus older children, making it impossible to study the full trajectory of vocal development. That is, qualitative shifts like the onset of canonical transitions may provide useful developmental markers. But once this onset has occurred, they do not provide further information on the child, whereas quantitatively varying

measures can. In fact, a recent paper suggests that canonical proportion continues increasing beyond two years of age (Hitczenko et al., 2023).

Here, we present a new, publicly-available dataset that will allow us to significantly expand our knowledge of early vocal development. This dataset consists of child vocalizations taken from naturalistic long-form recordings, labeled with their vocalization type: laughing, crying, canonical (speech-like vocalization containing an adjacent consonant and vowel), or non-canonical (speech-like vocalization without an adjacent consonant and vowel). These labels were chosen because they constitute the main vocalization types that children produce (Buder, Warlaumont, & Oller, 2013; Kent, 2022; Morgan & Wren, 2018), allowing researchers to study key indices of vocal development (e.g., *linguistic proportion*, the percentage of child vocalizations that are speech-like, and *canonical proportion*, the percentage of speech-like vocalizations that combine a consonant and vowel). Thanks to citizen science-based crowdsourcing, the dataset is also unprecedented in size, consisting of 256,842 clips from 382 children from 10 different communities (Table 1), and can be accessed at https://osf.io/tf3hq. Our corpus includes and builds on a previous smaller one called BabbleCor (Cychosz et al., 2021; used in the Interspeech 2019 ComParE challenge; Schuller et al., 2019). BabbleCor contained ~15,000 child vocalizations from 52 children, aged 1-36mo, growing up in a variety of cultural and linguistic environments. We go well beyond BabbleCor by providing labeled vocalizations from ~400 new children growing up in 4 new environments and adding over 15 times the amount of original data. In addition, our corpus contains labeled vocalizations both from the key child (i.e., the child wearing the audio recorder) and other speakers in their surroundings, which will

allow researchers to not only study how children's vocal development proceeds, but also how it relates to the linguistic environment they grow up in.

Our dataset addresses the three limitations noted above by drawing from naturalistic recordings. This method allows a less skewed participant sample: Our corpus includes children learning 41 different languages, growing up across 4 continents in both urban and rural, small-scale, subsistence-level environments, as well as both monolingual and multilingual settings, heeding calls for the diversification of participant samples (Singh et al., 2023) and providing a more substantive base for generalization. We particularly highlight the inclusion of multilingual participants which to our knowledge have very seldom been studied (Meera, Swaminathan, Ranjani V, Srikar, & Raju, 2023). Further, the long-form technology we use allows us to non-intrusively collect hours of data for each child, and study child and adult vocalizations as they happen in everyday language learning environments, offering a perspective that by definition reflects children's everyday productions. Finally, while past research has primarily focused on children aged 6 to 18 months old, our sample includes children as young as 1 month and through 6 years of age, covering the whole period during which previous work suggests there are quantitative shifts in canonical proportion (Hitczenko et al., 2023). This allows us to study the full trajectory of vocal development, and observe continuous changes over the course of this time period, in addition to qualitative shifts (e.g., the onset of canonical transitions), which may be easier to notice in lab studies, but cannot, on their own, provide a complete understanding of vocal development.

Additionally, our dataset can be used as training data to develop automated systems that categorize vocalization types from long-form recordings (e.g., Al Futaisi, Zhang, Cristia,

Warlaumont, & Schuller, 2019; Li, Hasegawa-Johnson, McElwain, 2024; Zhang, Cristia, Warlaumont, Schuller, 2018). This would allow other researchers to apply these methods to their own audio recordings, significantly expanding how and where vocal development can be studied. There is a scarcity of child-centered, well-annotated, public data, which has held back the development of important tools, with obvious consequences for performance. For example, the DIHARD Challenge, a shared automatic speech recognition task on speaker diarization (i.e., labeling who spoke when) in challenging corpora, included child-centered data in the first two editions (Ryant et al., 2018, 2019). Child-centered data tended to lead to the lowest diarization performance of submitted models, reflecting the poor adaptation of our speech technology tools to the settings in which children actually learn language. In later editions, the child-centered dataset was removed (Ryant et al., 2021), because of licensing concerns (and, in our opinion, potentially also to avoid this particularly hard data set). The relatively poor performance of even basic diarization tools reflects, in our view, a paucity of public in-domain data, which can allow researchers to train and/or fine-tune their models. While keeping in mind ethical and legal considerations and by harnessing the power of citizen science, we have built a corpus that can be used to develop speech technology tailored for children in real world language learning environments.

2 Methods for Corpus Creation

See Figure 1 for an overview of the procedures we used to create this dataset, each described in more detail in the sections that follow.

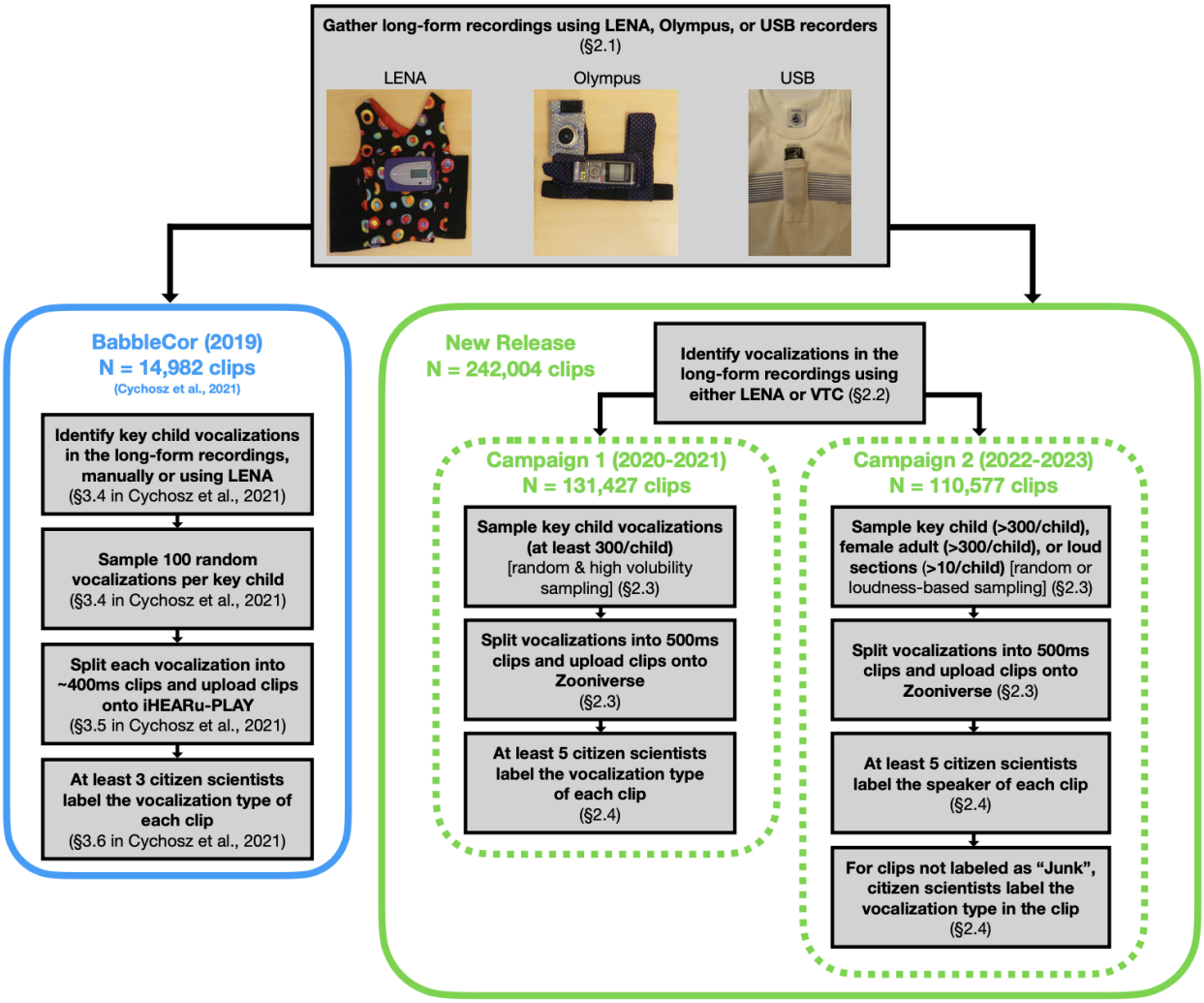

Figure 1. Flowchart depicting the creation of the dataset. The blue bubble corresponds to BabbleCor, a smaller, previously-created corpus that we build on here (and that is included in our current release). The solid green bubble corresponds to the new data in our dataset (this data was collected in two campaigns, depicted in dotted green bubbles). Each step of the process (shown in a gray box) contains a reference to a section in either this paper (for the new release) or Cychosz et al. (2021) (for BabbleCor), which describes the procedure in more detail.

2.1. Long-form recordings

CORPUS OF NATURALISTIC VOCALIZATIONS

The clips come from extant long-form recording corpora sampling from ten different communities worldwide and documented in previous work: Bolivia (Tsimane'; Scaff, Stieglitz, & Cristia, 2019; Scaff et al., 2024), Bolivia (Quechua & Spanish; Cychosz, 2018), France (French; Cristia, 2021), Mexico (Tseltal; Casillas, Brown, & Levinson, 2017; Casillas, Brown, & Levinson, 2020), Papua New Guinea (Yélî Dnye; Casillas, Brown, & Levinson, 2021; Cristia & Casillas, 2019), USA-Indiana (English; Hamrick, Seidl, & Kelleher, 2023; Semenzin, Hamrick, Seidl, Kelleher, & Cristia, 2021), USA-New York (English; Bergelson, 2017; Bergelson, Amatuni, Dailey, Koorathota, & Tor, 2018), USA-California (English & Spanish; Warlaumont, Pretzer, Mendoza, & Walle, 2016), Vanuatu (23 total languages are represented; each child was learning 1-8 of them; Cristia & Colleran, 2018), and Solomon Islands (12 total languages are represented; each child was learning at least 2 of them; Cassar, Cristia, Grosjean, & Walker, 2021). All children were typically-developing according to parental report, with the exception of N=10 children from the USA-Indiana (English) corpus who had severe language and developmental delays (Hamrick et al., 2023). Across the communities, children (aged 1mo-6yrs) were equipped with a non-intrusive audio recorder that they wore in a special shirt/vest pocket as they went about a typical day (see Figure 1). The audio recordings were either made using the Language ENvironment Analysis Digital Language Processor recording device (i.e., LENA; Ganek & Eriks-Brophy, 2018), or analogous Olympus or USB recorders (see Table 1 for details). Long-form recordings are an increasingly popular approach to study child development, as they can provide a glimpse into naturalistic learning environments across diverse communities and enable standardized collection of large amounts of data across them, including in communities that are not close to a traditional university-based infant lab (Lavechin, Seyssel, Gautheron, Dupoux, & Cristia, 2022).

In what follows, we describe the procedures for extracting and annotating vocalizations from these long-form recordings. We only describe the procedures for new clips not included in BabbleCor. The BabbleCor clips/annotations were collected in a very similar way, and we point the reader to Cychosz et al. (2021) and Figure 1 for those precise details.

2.2. Vocalization identification

We use automatic methods to identify *any* vocalizations within the speech recordings and categorize which speaker produced them as one of the following: key child (the child wearing the recorder), other child, adult female, or adult male. For LENA recordings, we do so using LENA's built-in speaker diarization algorithm. For non-LENA recordings, we do so using Voice Type Classifier (VTC), an open-source model specifically trained to identify vocalizations and their speakers from long-form recordings (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020). VTC and LENA identify key child and female adult vocalizations with relatively high reliability (VTC F-scores of 77% and 82%, respectively; LENA F-scores of 55% and 42%, respectively). Both algorithms are less reliable for male adult vocalizations, which also tend to be rare in child-centered recordings (VTC F-score = 42%; LENA F-score = 31%), so we do not rely on male adult speaker labels in the creation of this dataset (Cristia et al., 2021; Lavechin et al., 2020). Note that for the USA-Indiana (English) corpus, three lab-based annotators verified that key child speaker labels were correct.

2.3. Sampling vocalizations to be annotated

Annotating all of the identified vocalizations would be time-prohibitive (according to our rough estimates, labeling every vocalization could take 25+ years)[1], so we sampled a subset of vocalizations to be uploaded onto a citizen science platform for annotation by citizen scientists (as will be described in more detail in section 2.4). The precise number of vocalizations sampled varied as a function of different goals and research studies (see Table 1 for details). We used four primary approaches for sampling vocalizations across the corpora:

- **Key child (random) sampling**: In this approach, we randomly sampled segments automatically labeled as key child vocalizations. This approach gives a general view of children's vocal development.
- **Female adult (random) sampling**: In this approach, we randomly sampled segments automatically labeled as female adult vocalizations. This approach gives a view of the linguistic environment that children are learning in, specifically what types of

[1] As an example, children in the Solomon Islands corpus vocalize on average 8 minutes per hour of recording time. As will be described, we split all vocalizations into 500ms clips to protect the participants' privacy, which would result in ~1000 clips to be labeled per hour of recording time, or ~1000 clips x ~10 hours = ~10,000 clips to be labeled per long-form recording. Assuming roughly similar vocalization quantities across all ~N=400 children in the corpus, that would yield a total of 4,000,000 clips, which would require 20,000,000 individual classifications, as citizen science platforms recommend each clip be classified by at least 5 citizen scientists. At an average annotation rate of 2,000 classifications per day, this would take over 25 years.

vocalizations they hear. Because the automated methods have worse reliability in identifying male adult and other child vocalizations and because adult male vocalizations are rare in our recordings, we only sampled from female adults. We included this sampling approach for two of the corpora (for purely historical reasons): namely, the Vanuatu and Papua New Guinea (Yélî Dnye) corpora.

- **Loudness-based sampling**: The third approach was qualitatively different from the first two because we randomly sampled from the recordings based on intensity (i.e., perceived loudness), instead of sampling based on automated speaker labels. We identified the 5s-long windows of each recording that were 75th percentile or higher in terms of energy (dB SPL; only considering energy within the 50-3000Hz frequency band) and then randomly sampled ten of those loud windows per child. The 75th percentile threshold was established by testing a few different values of this parameter and comparing predictions against human-annotated sections. The reasoning for using this approach was to avoid any algorithm-driven biases, and to see whether this method could be as effective as the more resource-consuming and complicated algorithm-based one (preliminary results in Section 3.1 suggest it was not, as it resulted in more 'Junk' clips than the other sampling methods we use).

- **Key child (high volubility) sampling**: Finally, for one corpus only (the USA-Indiana corpus, and only a subset of its clips), we randomly sampled key child vocalizations from the portions of the recording with the highest child volubility (in addition to randomly sampling other vocalizations from the full recording, using the key child (random) sampling approach described above). The decision to include this sampling method was

made in the context of a study testing the reliability of the citizen science methods we use here (Semenzin et al., 2021). Including a variety of sampling methods allowed the researchers to assess the generalizability of their results across sampling methods.

To protect the privacy of the recorded families, we split each sampled vocalization into ~500ms clips and uploaded those clips onto the citizen science platform, instead of the full vocalizations. This approach keeps private information (e.g., names, addresses, etc.) unidentifiable, without compromising inter-annotator agreement (Semenzin et al., 2021). We uploaded these clips for annotation in two campaigns: Campaign 1 (2020-2021) consisted of 131,427 clips and Campaign 2 (2022-2023) consisted of 110,577 new clips (there was no overlap between Campaign 1 and Campaign 2 in terms of clips, although there was some overlap in terms of children). Campaign 1 only included key child vocalizations, while Campaign 2 included key child vocalizations, but also female adult vocalizations and vocalizations identified via loudness-based sampling; see Table 1 for more information about what was included in each campaign.

CORPUS OF NATURALISTIC VOCALIZATIONS

Table 1: Overview of the data included in the corpus. The table includes data from BabbleCor, as well as more recent Zooniverse campaigns, which are cued by the superscript ([0]BabbleCor; [1]Campaign 1; [2]Campaign 2).

Corpus	Total number children	Age range	Recorder type (Speaker identification method)	Vocalization sampling method [child details, if different from overall]	Number of vocalizations sampled	Total number 500ms clips	Annotated labels
Bolivia[0] (Quechua & Spanish)	N=3	24 mos	LENA (LENA)	Key child (random)	N=100/child	1,018	vocalization type
Bolivia[0,1] (Tsimane')	N=41	5-70 mos	LENA (LENA, VTC) & USB or Olympus (VTC)	Key child (random)	N=100-954/child	88,905	vocalization type
France[1] (French)	N=10	10-11 mos	LENA (LENA)	Key child (random)	N=300/child	3,476	vocalization type
Mexico[0] (Tseltal)	N=10	1-36 mos	Olympus (Manual)	Key child (random)	N=100/child	2,433	vocalization type
Papua New Guinea[0,1,2] (Yélî Dnye)	N=46	2-76 mos	LENA (LENA, VTC) & Olympus (Manual)	Key child (random) [N=46; 2-76 mos]	N=100-300/child	12,858	vocalization type
				Female adult (random) [N=5; key child age: 25-66 mos]	N=300/child	6,141	vocalization type & speaker
Solomon Islands[^,2]	N=199	4-48 mos	USB (VTC)	Loudness	N=10/child	11,000	vocalization type & speaker
USA-California[0] (English & Spanish)	N=3	3 mos	LENA (LENA)	Key child (random)	N=100/child	1,075	vocalization type

CORPUS OF NATURALISTIC VOCALIZATIONS

Corpus	Total number children	Age range	Recorder type (Speaker identification method)	Vocalization sampling method [child details, if different from overall]	Number of vocalizations sampled	Total number 500ms clips	Annotated labels
USA-Indiana[1] (English)	N=20	4-53 mos	LENA (LENA + Manual)	Key child (random) [N=20; 4-53 mos]	N=343-1038/child	10,569	vocalization type
				Key child (high volubility) [N=20; 4-53 mos]	N=343-1038/child	23,161	vocalization type
USA-New York[0] (English)	N=10	7-17 mos	LENA (LENA)	Key child (random)	N=100/child	2,914	vocalization type
Vanuatu*,[2]	N=40	5-51 mos	USB (VTC)	Female adult (random) [N=40; key child age: 5-51 mos]	N=300/child	41,606	vocalization type & speaker
				Loudness [N=40; 5-51 mos]	N=10/child	4,000	vocalization type & speaker
				Key child (random) [N=40; 5-51 mos]	N=300/child	47,830	vocalization type & speaker

**Languages represented in the Vanuatu corpus: Bislama, Venen Taut, Petarmul, Neverver, Uripiv, Vinmavis, Novol, Epi, Nah'ai, Paama, Ninde, Tautu, French, Pinalum, Malo, Rano, Tauta, Santo Language, Ambae, Maevo, South, Atchin, and Tempun. ^Languages represented in the Solomon Islands corpus: Roviana, Avaso, Babatana, Marco, Marovo, Pidjin, Senga, Simbo, Sisinga, Ughele, Vaghua, and Varisi*

2.4. Labeling the clips on Zooniverse

In Campaigns 1 and 2, after clips were identified and sampled by the process described in Sections 2.2-2.3, they were annotated by citizen scientists, as part of the Maturity of Baby Sounds project (https://www.zooniverse.org/projects/laac-lscp/maturity-of-baby-sounds) on Zooniverse (an online citizen science platform). Citizen science is a growing approach, in which volunteers, who need not have a specialized background in the area, help researchers in various aspects of their research. In our case, we asked citizen scientists to help label the clips we uploaded onto Zooniverse as follows:

- **Campaign 1 (2020-2021)**: In the first campaign (N = 131,427 clips), citizen scientists labeled the clips as one of the following (see Figure 2): laughing, crying, canonical (a speech-like vocalization that consisted of an adjacent vowel and consonant; “ma”, “am”), non-canonical (a speech-like vocalization that was only vowel or only consonant; “ahhh”, “gggh”), or junk (a clip that did not include a vocalization, or included overlapping speech/noise making it difficult to identify the properties of the vocalization). Prior to annotation, citizen scientists were provided with a tutorial that included explanations and examples of these vocalization types, which they could access at any time, and could play the sound as many times as they liked. In addition, they could reach out to the research team with any clarifications.

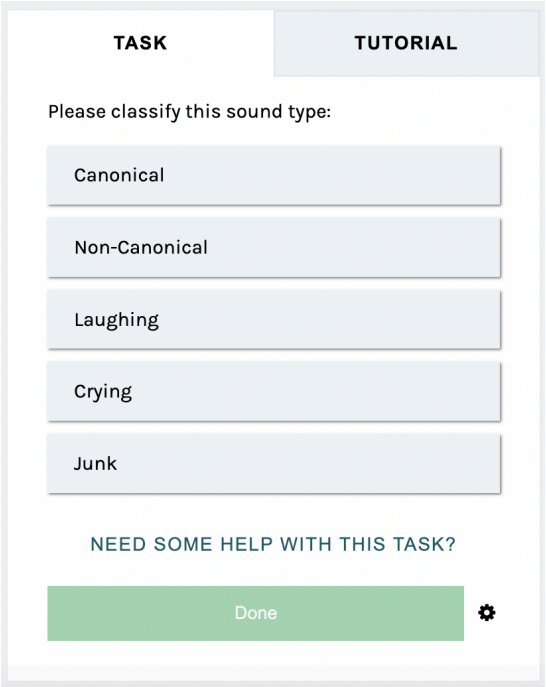

Figure 2. Zooniverse interface for labeling clips in Campaign 1. Citizen scientists labeled each clip as belonging to one of five categories, before moving onto the next clip. While BabbleCor used the iHEARu-PLAY platform instead of Zooniverse, the question was the same.

- **Campaign 2 (2022-2023)**: Because this Campaign included vocalizations by female adults and a sampling method in which the speaker is not known (loudness-based sampling), we added a couple of questions (Figure 3). In particular, citizen scientists first identified who they thought was vocalizing: a baby (defined as 0-3 years old), child (3-12 years old), adolescent (12-18 years old), adult (18+ years old), or if the clip fell into the 'Junk' category[2]. For clips that they identified as adolescent or adult speech, they also

[2] We proposed these categories as conceptually separable and potentially intuitive, but at the time, we couldn't be sure if listeners would be able to enforce it. After data were collected, we evaluated the proportion baby/(baby+child) among key child vocalizations as a function of the

labeled their perceived gender.[3] Finally, they were asked to classify the vocalization type (crying, laughing, canonical, non-canonical) for all non-junk clips. If the clip included multiple speakers, citizen scientists were asked to answer for the most prominent speaker in the clip, or to label the clip as "Junk" if overlapping speech made it impossible to identify the characteristics of the individual voices.

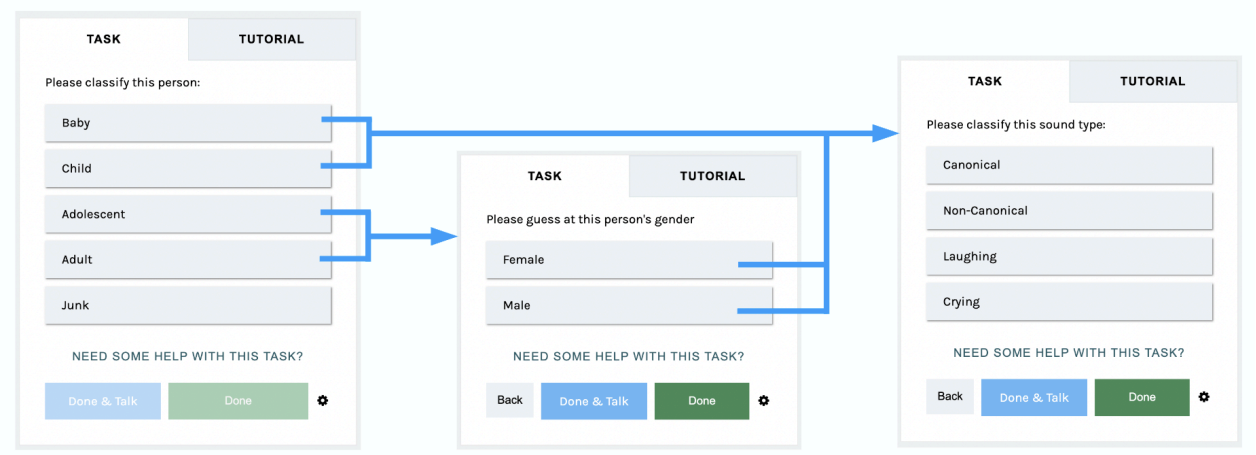

Figure 3. Zooniverse interface for labeling clips in Campaign 2. In Campaign 2, citizen scientists first labeled speaker information before deciding on vocalization type. The blue arrows indicate

---

real age of the recorded child. We found a clear discontinuity, with a clear drop in the proportion of clips recognized as "baby" by listeners when the recorded child was on average 2.5 years of age. This suggests to us that listeners are quite good at distinguishing between these categories, which is useful for cases in which the speaker's age is not known (i.e., when sampling vocalizations that are not by the key child).

[3] We acknowledge the inadequacy of reducing gender to a binary classification based solely on the sound of adolescent and adult voices in 500ms clips.

the flow of decisions, which was contingent upon the answers provided (e.g., baby/child vocalizations were not labeled for gender).

Following Zooniverse recommendations and systematic experimentation by Cychosz et al. (2021), each clip was labeled by at least three citizen scientists and the vast majority were labeled by five (Campaign 1: 93%; Campaign 2: 98%). The final labels for each clip were chosen based on "majority vote": if at least 50% of the citizen scientists endorsed a particular label (and no other label received the same number of votes), the clip received that label; otherwise it received a NO-LABEL label. For Campaign 2, this was done question by question, meaning that some clips may have e.g. speaker majority-labels, but no vocalization type majority-labels, if citizen scientists agreed on speaker, but not vocalization type. Note that even though each clip was labeled by at least three citizen scientists, some individual majority-labels in Campaign 2 could be decided based on fewer than three votes: for example, gender information, if one of three citizen scientists chose the 'child' label instead of 'adult' (they would not be prompted for gender information then), or vocalization type information, if one of three citizen scientists chose the 'Junk' label. See Table 2 for labeling examples. We release all individual labels in addition to majority labels, so users can choose alternative approaches in their analyses.

Table 2: Hypothetical examples of how final labels are decided based on 3 or more labels provided by individual annotators. In Campaign 2, there are three rows because annotators could provide labels for up to 3 questions (speaker, gender, vocalization type).

Clip	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Final Label
			BabbleCor + Campaign 1 of New Release			
1	Canonical	Canonical	Junk	-	-	Canonical
2	Crying	Crying	Canonical	Laughing	-	Crying
3	Crying	Crying	Canonical	Canonical	-	NO-LABEL
4	Junk	Junk	Junk	Canonical	Laughing	Junk
			Campaign 2 of New Release			
5	Adult Female Canonical	Adult Female Canonical	Child - Crying	Adult Female Canonical	Child - Crying	Adult Female Canonical
6	Adult Female Canonical	Adult Male Laughing	Child - Canonical	Adult Female Laughing	Adult Male Canonical	Adult NO-LABEL Canonical
7	Junk - -	Adult Male Canonical	Baby - Canonical	Junk - -	Adult Female Laughing	NO-LABEL NO-LABEL Canonical
8	Adult Female Canonical	Adult Male Laughing	Adult Female Canonical	Adolescent Male Crying	Junk - -	Adult NO-LABEL Canonical
9	Junk - -	Junk - -	Adult Female Canonical	Adolescent Female Canonical	Junk - -	Junk Junk Junk

3 Dataset analyses

3.1. Descriptive statistics about annotators and annotations

In total, over 26,673 unique annotators provided 1,194,870 total labels on the 242,004 clips we uploaded to Zooniverse. Each citizen scientist labeled an average of 45 and a median of 2 clips (range 1-44604).

Across Campaigns 1 and 2, 72,174 clips (30%) did not receive a majority vocalization label. Of the remaining 169,830 clips (which did receive a majority vocalization label), 73,979 were labeled as Junk (44%), 28,356 as Canonical (17%), 47,620 as Non-Canonical (28%), 4,360 as Laughing (3%), and 15,515 as Crying (9%) (see Figure 4 for by-corpus numbers). For the same information on BabbleCor, please refer to Cychosz et al. (2021).

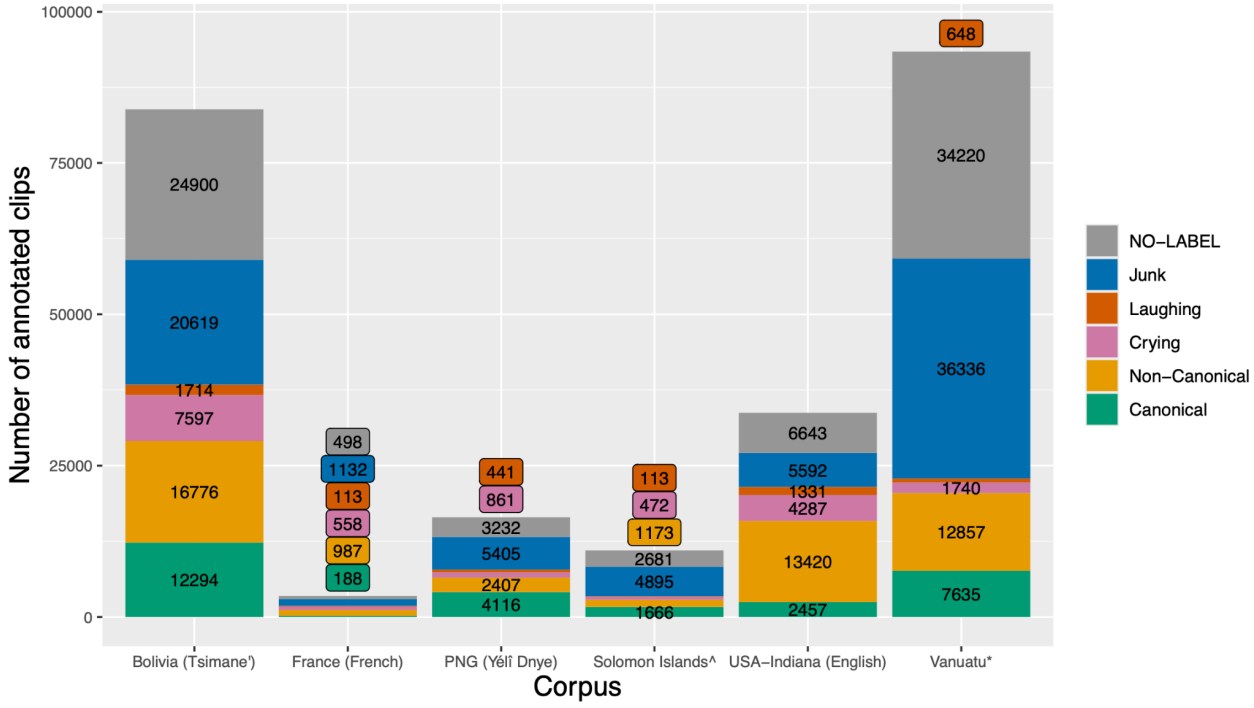

Figure 4. Vocalization type annotations by corpus included in Zooniverse campaigns. We place labels that would lead to textual overlap above the bars.

In Campaign 2 (N=110,577), in which citizen scientists additionally labeled speaker information, 36,937 total clips (33%) did not receive a majority label for age (i.e., baby, child, adolescent, adult). Of the remaining 73,640 clips (which did receive a majority age label), 17,741 were labeled as babies (24%), 6,957 as children (9%), 853 as adolescents (1%), 4,686 as adults (6%), and 43,403 as junk (59%). We note that differentiating some of the speaker groups (e.g., babies vs. children or adolescents vs. adults) may be difficult on the basis of ~500ms clips, so we would expect higher agreement if collapsing across these categories in analyses. We provide raw classifications, which allows dataset users to recalculate majorities as they would like.

Of the 5,539 clips that were majority-labeled as adolescent/adult (i.e., those for which citizen scientists were also asked to provide gender information), 1,183 total clips (21%) did not

receive majority labels for gender. Of those that did, 2,951 were labeled as Female (68%) and 1,405 as Male (32%).

In sum, combined with BabbleCor, the full corpus consists of 30,184 clips labeled Canonical, 53,873 clips labeled Non-Canonical, 16,291 clips labeled Crying, 4,535 clips labeled Laughter, and 79,191 clips labeled Junk. In terms of speaker, the full corpus consists of 24,698 clips labeled as Baby/Child (with an additional 23,161 clips that were not labeled for speaker, but were randomly sampled from vocalizations that were automatically identified to be the key child) and 5,539 clips labeled as Adolescents/Adults (2,951 female, 1,405 male, and 1,183 without a gender label).

3.2. Interannotator agreement levels

Each clip in the corpus was labeled by at least 3 citizen scientists (and most were labeled by 5 or more). In this section, we ask what the level of interannotator agreement was. Figure 5 shows the agreement rates by campaign (i.e., BabbleCor vs. Zooniverse) while Figure 6 shows the agreement rates by corpus. Overall, we observe lower agreement rates for the Zooniverse data than for the BabbleCor data (which had very high agreement rates). For example, the proportion of clips that did not receive a majority-label (i.e., clips on which citizen scientists do not agree) is higher on Zooniverse than what was observed in BabbleCor (~30% vs. 5%; Figure 5). Potentially related, we found that there were relatively more Junk labels in the Zooniverse data than on the BabbleCor data.

Some of the difference in agreement rates is likely due to the fact that Zooniverse clips were labeled by more citizen scientists than BabbleCor clips (most Zooniverse clips received 5 or

more labels, while most BabbleCor clips received 3), which makes it harder to achieve a majority. To illustrate, if we assume annotators choose between the five options at random, the probability of achieving a majority purely by chance is 0.52 with three annotators (majority >= 2/3), but only 0.29 with five annotators (majority >= 3/5). For this reason, we caution against reading too much into the precise difference in agreement rates, as it likely overstates the true difference between campaigns.

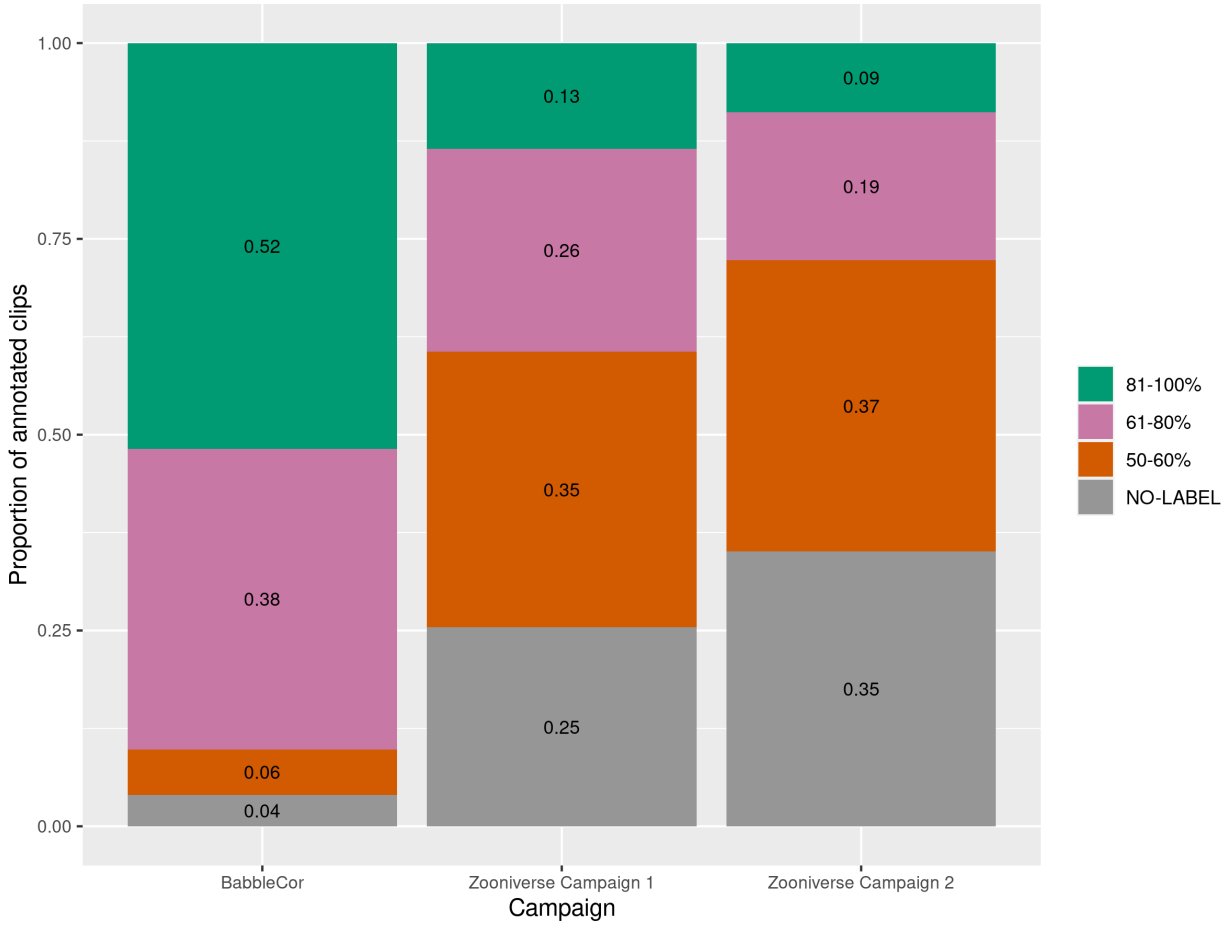

Figure 5. Interannotator agreement rates by campaign (BabbleCor and two Zooniverse campaigns). Most BabbleCor clips were labeled by 3 citizen scientists, while most Zooniverse

clips were labeled by 5 citizen scientists. Overall, we observe higher agreement rates on BabbleCor than on the Zooniverse campaigns.

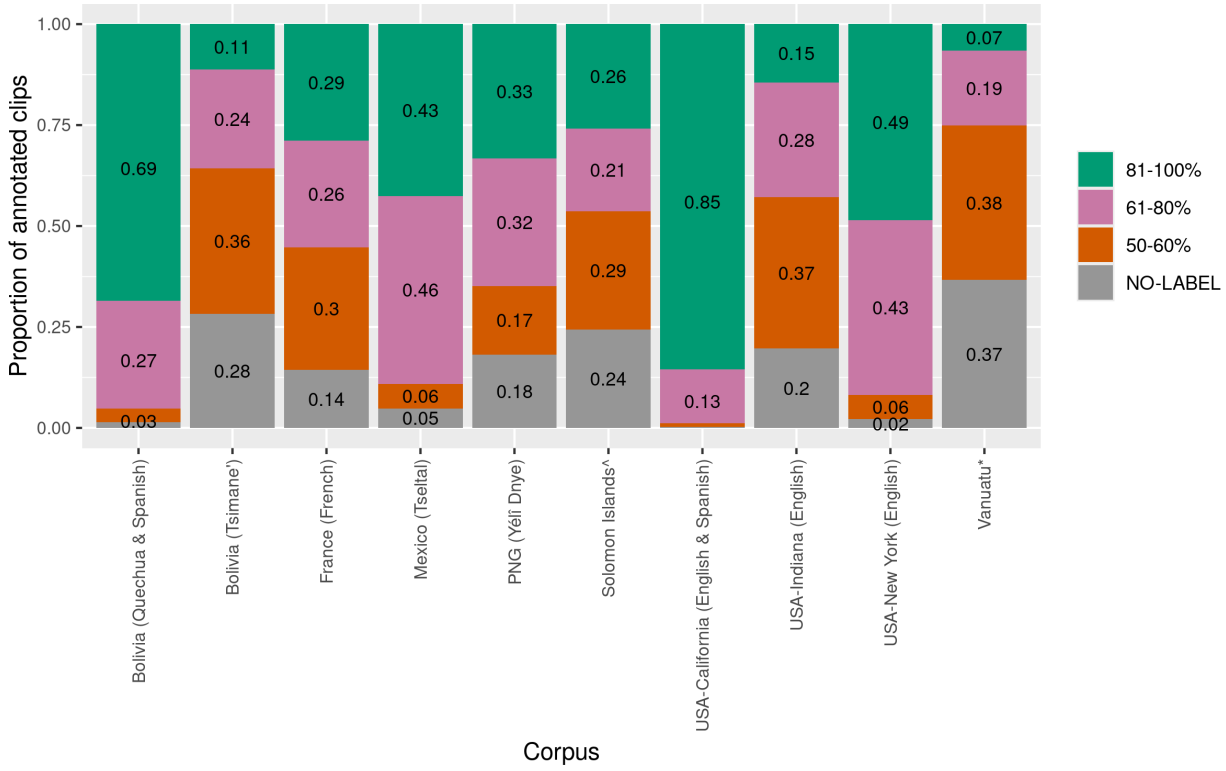

Figure 6. Interannotator agreement rates by corpus. We observe substantial variability between corpora, most likely due to the equipment/algorithms used to analyze them, rather than inherent properties of the corpora (e.g., language spoken).

That said, given how large the difference in agreement rates between campaigns is, there are likely a number of additional factors that contribute to this difference between campaigns, and we think the three most likely are (1) the equipment and/or algorithm used for the uploaded clips,

(2) differences in annotator training (experience and motivation) across the two platforms, and (3) differences between the samples included in each campaign, which we discuss in turn.

First, the four rightmost columns of Figure 7 show higher agreement rates on clips that were identified manually and/or collected using LENA devices and analyzed using LENA software than on clips that were collected using other devices (i.e., Olympus or USB devices) and identified using VTC (we plot data from Tsimane', different subsets of which were collected using different approaches and, thus, allows for a clear comparison). BabbleCor consisted exclusively of hand-annotated or LENA-derived data, while the more recent data collection campaigns, labeled through Zooniverse, also included VTC clips. Higher agreement with LENA than VTC could be due to two differences across these algorithms: a different compromise between precision and recall, and the possibility of overlap. As to the former reason, LENA's key child label has a higher precision (67% for the key child on a reference corpus[4]) at the expense of a lower recall (47% respectively), meaning that the algorithm is conservative by only calling a section key child vocalizations when it is quite certain. VTC is more balanced, sacrificing some precision (57% for the key child on the same reference corpus) in order to capture a higher

[4] The reference corpus was based on human annotation performed on 7 corpora available to our team, only partially overlapping with those sampled for the present paper. The corpus was quite large (based on tens of hours of human annotations) and diverse (including data from English, French, Tsimane', and Yélî Dnye). We presume also that it is representative of the way VTC and LENA differ in terms of precision/recall trade-off. Documentation for this analysis can be found in

https://gin.g-node.org/LAAC-LSCP/reliability/src/master/output/aggregate_recall_precision.csv

proportion of key child vocalizations, resulting in a higher recall (71%). As to overlap, LENA classifies key child vocalizations that overlap with other voices and/or with noise as "overlap" (and they lose their "key child" label), whereas VTC allows vocalizations to overlap, so they retain their "key child" label. This means that one cannot sample such vocalizations using LENA-derived clips, whereas one can when using VTC. As a result of these two differences, VTC may return a higher number of key child vocalizations, some of which are actually not the key child (due to lower precision) and some that have overlapping speech or noise, which will lead to Junk or NO-LABEL because it is harder to make out who is speaking and because citizen scientists were instructed to label overlapping speech as "Junk" if it was impossible to discern the individual voices.

That said, the two leftmost columns in Figure 7 only consider LENA data, but still show higher agreement rates in the BabbleCor vs. Zooniverse annotation campaign, suggesting that differences in equipment/algorithms do not tell the full picture. A second likely explanation for this difference is that the BabbleCor platform had fewer annotators (often, students of the researchers involved in BabbleCor, or the researchers themselves) and all annotators underwent specific training for the task (including a quiz that tested their understanding of distinctions like canonical vs. non-canonical). In contrast, the Zooniverse platform was open to the general public and, while annotators were provided with detailed instructions and examples of different vocalization types (e.g., canonical vs non-canonical), the training was less extensive. As a result, there were differences in the number of annotations per annotator across the two platforms. BabbleCor had 136 unique annotators (for 14,982 clips) while Zooniverse had 26,673 unique annotators (for 242,004 clips). The average BabbleCor citizen scientist annotated 478 clips (vs.

45 clips on Zooniverse) and, thus, gained more experience. We think that more training/experience may have led to higher agreement rates on BabbleCor.

Finally, there were qualitative differences between the type of data included in the three campaigns, including age, sampling, and cultural/environmental differences, which could have also affected interannotator agreement and Junk rates. First, the key children whose audio recordings were included in BabbleCor were younger than those included in the Zooniverse campaigns, and it is possible that annotators have an easier time annotating the shorter, less complex vocalizations that younger infants produce. Second, the clips included in BabbleCor only used the key child random sampling approach, while clips included in Zooniverse were also sampled using three other approaches (female adult random sampling, loudness-based sampling, and key child high volubility), which likely increased NO-LABEL and Junk rates. Specifically, since the loudness-based approach is based on intensity and not algorithm-derived speaker labels, this approach is likely to include many clips that do not contain human vocalizations (e.g., animal sounds) and may oversample overlapping speech, which result in higher Junk rates. Similarly, because the key child wears the audio recorder, sampling from female adult speech may also increase Junk and NO-LABEL rates, as the presence of adult speech necessarily means there are multiple speakers present (which could lead to more overlapping speech), and the adult speech is farther from the audio recorder than key child speech, which may make it harder to classify. Finally, the communities included in each campaign differed systematically. For example, the fact that we observe more “Junk” in the Zooniverse is, at least partially, likely because larger proportions of the Zooniverse data came from communities in which children are more likely to spend time outside, with potentially more background noises (e.g., animal sounds), and around a

greater number of speakers. In sum, these differences in sampling could contribute to higher disagreement (NO-LABEL) and Junk rates, in addition to the other methodological differences discussed.

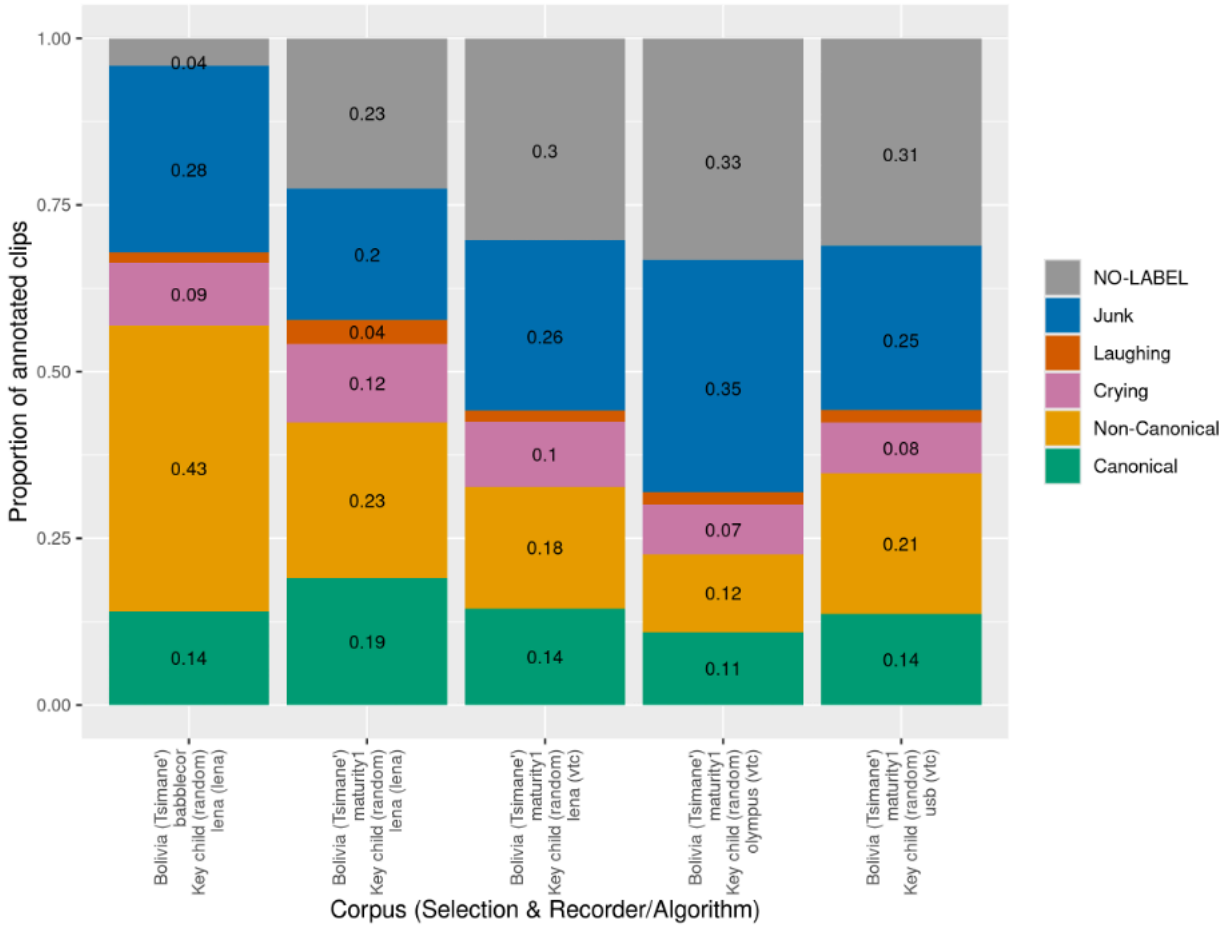

Figure 7. Vocalization type annotation rates for the Tsimane' data, which was collected and analyzed using a variety of different recorder types and algorithms. Although the collected data are different across the columns (e.g., Olympus in Bolivia was used only with 2 children under 2 years of age), this nonetheless allows for a direct comparison, which clearly shows that NO-LABEL and Junk rates are lowest for data collected and analyzed using LENA (though differences in age, sampling approach, etc. also likely contribute).

Overall, excluding Junk and NO-LABELs, our corpus provides over 104,883 vocalization clips with high-quality, manual citizen science annotations, representing an over ten-fold increase over BabbleCor (which included 9,032 clips with majority agreement). Next, we demonstrate with a case example how our corpus can be used to study language acquisition.

3.3. Case Study: Linguistic and Canonical Proportion as a Function of Speaker Age

As described in the Introduction, we envision many potential use cases for this corpus. Here, we use the corpus to study how canonical proportion (the proportion of speech-like vocalization clips that have a consonant-vowel/vowel-consonant transition) and linguistic proportion (the proportion of vocalization clips, including laughing and crying, that have a speech-like - canonical or non-canonical - vocalization) change as a function of developmental age (baby, child, adolescent, adult). We first exclude all clips that received NO-LABEL or Junk labels. We then calculate separate canonical proportions and linguistic proportions for each speaker type (baby, child, adolescent, adult) that appears in each key child's audio recording. To ensure that the proportions are based on a sufficient amount of data, we exclude proportions that are based on fewer than 20 audio clips per key child (i.e., if there are only 3 audio clips labeled as adolescent in a particular key child's audio recording, we do not include an adolescent canonical/linguistic proportion for that key child).

Figure 8 shows the canonical/linguistic proportions across speaker types - across all corpora (left), as well for the two corpora from which (for purely historical reasons) we also uploaded female, adult vocalizations (namely, the Vanuatu corpus - middle - and the Papua New Guinea corpus - right). For linguistic proportion (top row), we observe that the linguistic proportions of children, adolescents, and adults have all plateaued at nearly 100%, whereas

babies (younger than 3 years old) show lower linguistic proportions of around 80%. For canonical proportion, on the other hand, we see that only the adolescents and adults have plateaued at canonical proportions of 85-90%. Clips attributed to babies, and interestingly children (aged 3-12), have lower canonical proportions, suggesting that their canonical development is still ongoing. This matches with recent data from Hitczenko et al. (2023), but is at odds with a typical assumption in the field, that such global measures of vocal development are no longer of interest past the first-word stage (at around one year of age). Overall, these results provide further evidence that canonical proportion, a key measure of vocal/phonological development, continues to increase and should be studied even after children's vocalizations are thought to be driven by word choice, and show how this dataset can be used to gain insights into language development.

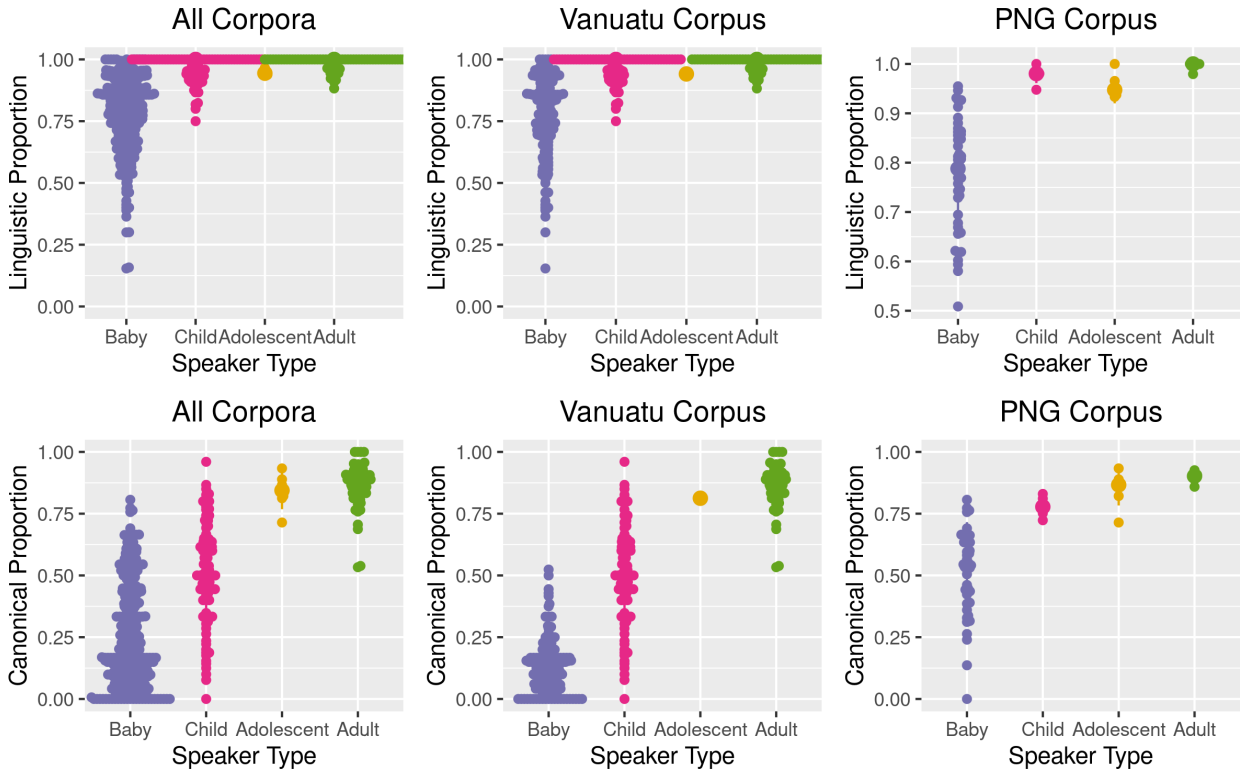

Figure 8. Linguistic proportions (top) and canonical proportions (bottom) by perceived speaker age (baby, child, adolescent, adult), calculated across all corpora included in the Speech Maturity Dataset (left column); only participants in the Vanuatu corpus (middle column), and only participants in the Papua New Guinea corpus (right column). Each point represents the data extracted from one key child's audio recordings (i.e., linguistic/canonical proportion calculated over all clips labeled as baby/child/adolescent/adult in one key child's audio recording). Linguistic proportion only shows significant variation in clips attributed to babies, whereas canonical proportion also shows variability in clips attributed to children.

4 Conclusion

In this paper, we have presented the Speech Maturity Dataset, a publicly-available corpus of children's vocalization development, including naturalistic data collected from over 400 children around the world. The diverse sample includes children aged 1mo-6yrs from 10 different communities, learning 25+ languages, in both urban, industrialized and rural, small-scale, subsistence-level environments, as well as both monolingual and multilingual settings. This dataset represents an ongoing and collaborative effort between field researchers, psycholinguists, computational linguists, and citizen scientists (overall, more than 20,000 individuals), and has the potential to transform the scale at which the field can study vocal/phonological development.

While we have provided one example of how this corpus can help us study vocal development, future approaches can more specifically delineate the trajectory of vocal development over a wide age range, study the extent to which vocal development proceeds similarly across children or potentially depends on ambient linguistic and environmental factors, perform acoustic analyses of children's vocalizations over developmental time, and much more. In addition, this dataset can be used to train machine learning algorithms that automatically label and extract key developmental measures from long-form recordings collected around the world, which can further facilitate the study of language development in understudied populations.

Overall, this corpus shows how combining automated approaches with citizen science approaches can be used to measure cross-linguistic language development in natural settings at an unprecedented scale. We hope that this corpus will both lead to new discoveries in the area of phonological/vocal development, as well as encourage similar efforts in other domains.

Declarations

Funding: This work was funded by 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). This publication uses data generated via the Zooniverse.org platform, the development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. The funders had no impact on this study.

Competing Interests: The authors have no competing interests to disclose.

Data Availability: The dataset can be accessed at: https://osf.io/tf3hq/.

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. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3340555.3353751

Bergelson, E. (2017). *Bergelson Seedlings HomeBank corpus*. https://doi.org/https://doi.org/10.21415/T5PK6D

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2018). 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., et al.others. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120.

Buder, E. H., Warlaumont, A. S., & Oller, D. K. (2013). An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective. In *Comprehensive perspectives on child speech development and disorders: Pathways from linguistic theory to clinical practice*. Hauppauge, NY: Nova Science Publishers, Inc.

Casillas, M., Brown, P., & Levinson, S. C. (2017). *Casillas HomeBank Corpus*. https://doi.org/doi:10.21415/T51X12

CORPUS OF NATURALISTIC VOCALIZATIONS

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early Language Experience in a Tseltal Mayan Village. *Child Development*, *91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. *Journal of Child Language*, *48*(4), 792–814. https://doi.org/10.1017/S0305000920000549

Cassar, A., Cristia, A., Grosjean, P., & Walker, S. (2021). *Long-form recordings in the Solomon Islands*.

Cristia, A. (2021). *PhonSES: A pilot study to measure socioeconomic status association with infants' word and sound processing*. Retrieved from https://gin.g-node.org/LAAC-LSCP/phonSES-public

Cristia, A., & Casillas, M. (2019). *LENA recordings in Rossel Island*.

Cristia, A., & Colleran, H. (2018). *Long-form, child-centered recordings collected in Malekula in 2016-2018.*

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., … Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

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

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., … Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*, *24*(5), e13090. https://doi.org/10.1111/desc.13090

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.

Hamrick, L. R., Seidl, A., & Kelleher, B. L. (2023). Semi-automatic assessment of vocalization quality for children with and without angelman syndrome. *American Journal on Intellectual and Developmental Disabilities*, *128*(6), 425–448.

Hitczenko, K., Bergelson, E., Casillas, M., Colleran, H., Cychosz, M., Grosjean, P., … Cristia, A. (2023). The development of canonical proportion continues through 6 years of age. *Proceedings of the 20th International Congress of Phonetic Sciences*. Prague, Czech Republic.

Kent, R. D. (2022). The maturational gradient of infant vocalizations: Developmental stages and functional modules. *Infant Behavior and Development*, *66*, 101682. https://doi.org/10.1016/j.infbeh.2021.101682

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). *An open-source voice type classifier for child-centered daylong recordings. Interspeech.* https://doi.org/10.48550/arXiv.2005.12656

Lavechin, M., Seyssel, M. de, Gautheron, L., Dupoux, E., & Cristia, A. (2022). Reverse Engineering Language Acquisition with Child-Centered Long-Form Recordings. *Annual Review of Linguistics*, *8*, 389–407.

Lewedag, V. L., Oller, D. K., & Lynch, M. P. (1994). Infants' vocalization patterns across home and laboratory environments. *First Language*, *14*(42-43), 049–065. https://doi.org/10.1177/014272379401404204

Li, J., Hasegawa-Johnson, M., & McElwain, N. L. (2024). Analysis of self-supervised speech models on children's speech and infant vocalizations. *IEEE ICASSP Workshop on Self-supervision in audio, speech, and beyond.* Seoul, South Korea.

Majorano, M., Vihman, M. M., & DePaolis, R. A. (2014). The Relationship Between Infants' Production Experience and Their Processing of Speech. *Language Learning and Development*, *10*(2), 179–204. https://doi.org/10.1080/15475441.2013.829740

Meera, S., Swaminathan, D., Ranjani V, S., Srikar, M., & Raju, R. (2023). Canonical babbling ratio extracted from day-long audio recordings: A preliminary report from India. *Proceedings of the 20th International Congress of Phonetic Sciences*, 1201–1205. Prague, Czech Republic.

Morgan, L., & Wren, Y. E. (2018). A systematic review of the literature on early vocalizations and babbling patterns in young children. *Communication Disorders Quarterly*, *40*(1), 3–14.

Oller, D. K. (2000). *The Emergence of the Speech Capacity*. New York: Psychology Press. https://doi.org/10.4324/9781410602565

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2018). *First DIHARD challenge evaluation plan* [Technical {Report}]. Retrieved from https://catalog.ldc.upenn.edu/docs/LDC2019S12/first_dihard_eval_plan_v1.3.pdf

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2019). The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines. *Interspeech 2019*, 978–982. ISCA. https://doi.org/10.21437/Interspeech.2019-1268

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., … Liberman, M. (2021). The Third DIHARD Diarization Challenge. *Interspeech 2021*, 3570–3574. ISCA. https://doi.org/10.21437/Interspeech.2021-1208

Scaff, C., Stieglitz, J., & Cristia, A. (2019). *Excerpts from daylong recordings of young children learning Tsimane' in Bolivia*. https://doi.org/DOI 10.17605/OSF.IO/5869Q

Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2024). Characterization of children's verbal input in a forager‐farmer population using long‐form audio recordings and diverse input definitions. Infancy, 29(2), 196-215. https://doi.org/10.1111/infa.12568

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A. S., Yankowitz, L., Noeth, E., Amiriparian, S., Hantke, S., & Schmitt, M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian dialects,

continuous sleepiness, baby sounds & orca activity. In G. Kubin, & Z. Kačič (Eds.), Interspeech (pp. 2378–2382). International Speech Communication Association.

Semenzin, C., Hamrick, L., Seidl, A., Kelleher, B. L., & Cristia, A. (2021). Describing Vocalizations in Young Children: A Big Data Approach Through Citizen Science Annotation. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2401–2416. https://doi.org/10.1044/2021_JSLHR-20-00661

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. M. (2023). Diversity and representation in infant research: Barriers and bridges toward a globalized science of infant development. *Infancy*, *28*(4), 708–737.

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

Yankowitz, L. D., Schultz, R. T., & Parish-Morris, J. (2019). Pre-and paralinguistic vocal production in ASD: Birth through school age. *Current Psychiatry Reports*, *21*, 1–22.

Zhang, Z., Cristia, A., Warlaumont, A. S., & Schuller, B. (2018). Automated classification of children's linguistic versus non-linguistic vocalisations. *Interspeech 2018*. Hyderabad, India.