BEHAVIOR RESEARCH METHODS

A thorough evaluation of the Language Environment Analysis (LENA) system

Journal:	*Behavior Research Methods*
Manuscript ID	BR-Org-19-300
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	02-Aug-2019
Complete List of Authors:	Cristia, Alejandrina; LSCP, D\'{e}partement d'\'{e}tudes cognitives, ENS, EHESS, CNRS, PSL Research University, Lavechin, Marvin; LSCP, D\'{e}partement d'\'{e}tudes cognitives, ENS, EHESS, CNRS, PSL Research University Scaff, Camila; LSCP, D\'{e}partement d'\'{e}tudes cognitives, ENS, EHESS, CNRS, PSL Research University Soderstrom, Melanie; University of Manitoba Rowland, Caroline; Max Planck Institute for Psycholinguistics Rasanen, Okko; Tampere University; Aalto University Bunce, John; University of Manitoba Bergelson, Elika; Duke University

A thorough evaluation of the Language Environment Analysis (LENA) system

Alejandrina Cristia[1], Marvin Lavechin[1], Camila Scaff[1], Melanie Soderstrom[2], Caroline Rowland[3], Okko Räsänen[4,5], John Bunce[2,], & Elika Bergelson[6]

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

[2] Department of Psychology, University of Manitoba, Canada

[3] Max Planck Institute for Psycholinguistics, Netherlands

[4] Unit of Computing Sciences, Tampere University, Finland

[5] Department of Signal Processing and Acoustics, Aalto University, Finland

[6] Psychology & Neuroscience, Duke University, Durham, North Carolina, USA

Author Note

Correspondence concerning this article should be addressed to Alejandrina Cristia, 29, rue d'Ulm, 75005 Paris, France. E-mail: alecristia@gmail.com

Abstract

In the previous decade, dozens of studies involving thousands of children across several research disciplines have made use of a combined daylong audio-recorder and automated algorithmic analysis called the LENA® system, which aims to assess children's language environment. While the system's prevalence in the language acquisition domain is steadily growing, there are only scattered validation efforts, on only some of its key characteristics. Here, we assess the LENA® system's accuracy across all of its key measures: speaker classification, adult word counts (AWC), child vocalization counts (CVC), and conversational turn counts (CTC). Our assessment is based on manual annotation of clips that have been randomly or periodically sampled out of daylong recordings, collected from (a) populations similar to the system's original training data (North American English-learning children aged 3-36 months), (b) children learning another dialect of English (UK), and (c) slightly older children growing up in a different linguistic and socio-cultural setting (Tsimane' learners in rural Bolivia). We find reasonably high accuracy in some measures (AWC, CTC), with more problematic levels of performance in others (CTC, precision and recall of male adults and other children). We find little difference in accuracy as a function of child age, dialect, or socio-cultural setting. Whether LENA® results are accurate enough for a given research, educational, or clinical application depends largely on the specifics at hand. We therefore conclude with a set of recommendations to help researchers make this determination for their goals.

Keywords: Speech technology; human transcription; English; Tsimane'; Reliability; Agreement; Method comparison; Measurement error; Child vocalization count; Adult word count; Conversational turn count; LENA

A thorough evaluation of the Language Environment Analysis (LENA) system

While nearly all humans eventually become competent users of their language(s), documenting the experiential context of early acquisition is crucial for both theoretical and applied reasons. Regarding theory, there are many open questions about what kinds of experiences and interactions are necessary, sufficient, or optimal for supporting language development. Moreover, the ability to accurately and quickly assess an infant's state of development at a given point in time is of central importance for clinical purposes, both for children with known risks of language delays and disorders, and those who might not be identified based on risk factors. Reliable assessments are also crucial for measuring intervention efficacy.

One approach that has been making its way into the mainstream literature across basic and applied research on language and cognition relies on day-long recordings gathered with a LENA® audiorecorder (Gilkerson et al., 2017; e.g., Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2011; Oller et al., 2010; VanDam & De Palma, 2018), and further analyzed using automated, closed-source algorithms. As we summarize below, this approach has many advantages, which may explain its expanding popularity. While over a hundred papers over the past two decades have used the output automatically provided by LENA®, only a handful include validity estimates (d'Apice, Latham, & Stumm, 2019; e.g., Weisleder & Fernald, 2013; Zimmerman et al., 2009), even fewer where validity estimation was the primary focus of the paper (Bulgarelli & Bergelson, 2019; e.g., Busch, Sangen, Vanpoucke, & Wieringen, 2018; Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2016; Ganek & Eriks-Brophy, 2018; Lehet, Arjmandi, Dilley, Roy, & Houston, 2018). As a result, few studies report sufficient details about validation accuracy for one or more metrics, limiting the interpretability of the results of a meta-analytic assessment (cf. A. Cristia, Bulgarelli, & Bergelson, under review). The work undertaken thus far also has some limitations, which are described further in the "Previous Validation"" section below, and mentioned briefly next. First, many validations or evaluations of LENA® omit analysis of the less-directly "relevant"

categories of input like noise, silence, or overlap. Second, many LENA® evaluations rely on the output from LENA® as a starting point to either select portions of the file for manual annotation, or use its segmentation into talker turns or vocalizations as the unit of analysis rather than segmenting the audio from scratch. Both decisions can lead to inflated accuracy estimates. Here we endeavor to conduct an evaluation that is fully independent of the LENA® algorithms' automated assessment, permitting a systematic, extensive, and independent evaluation of its key metrics.

In a nutshell, this paper reports on the performance of LENA® algorithms when compared to human annotations carried out in a set of clips extracted from daylong audiorecordings gathered from (a) a sample of children similar to the LENA® training set (i.e. infants and toddlers, growing up in North American English-speaking homes, and aged 3-36 months), (b) a group of similarly aged children learning a different dialect (UK English); and (c) slightly older children learning a different language in a very different sociocultural setting (Tsimane'-learning children in rural Bolivia).

Brief introduction to LENA® products. The LENA® system consists of hardware and software. The hardware component is a lightweight, sturdy, and easy-to-use recording device worn by a child in specialized clothing. The software is a suite of proprietary computer programs designed to provide automated quantitative analyses of the children's auditory environment and their own vocalizations. The latter was developed over an extensive corpus of full day audio recordings gathered using their patented recording hardware (D. Xu, Yapanel, & Gray, 2009). The original dataset included over 65,000 hours of recording across over 300 American English-speaking families chosen for diversity in child age (1-42 months) and socio-economic status (Gilkerson & Richards, 2008). Half-hour selections from 309 recordings were transcribed and annotated for the purpose of developing the algorithm, and an additional 60 minutes from 70 additional recordings were transcribed and annotated for testing the result (Gilkerson, Coulter, & Richards, 2008).

The resulting LENA® software takes as input a new audio recording and processes it

incrementally in short windows, extracting a variety of acoustic features which are used to classify the audio stream into segments of at least 600 ms in length (or longer for some of the categories) using a Minimum Duration Gaussian Mixture Model (MDGMM; D. Xu et al., 2009). Silence may be included to "pad" segments to this minimum duration. The segments are classified according to a set of broad speaker and non-speaker classes. The speaker classes are: Male Adult, Female Adult, "Key" Child (i.e. the one wearing the recorder) and Other Child. The non-speaker classes are: Noise, Television (including any electronics), Overlap (speech overlapped with other speech or nonspeech sounds), and Silence (SIL). With the exception of Silence, these classifications are then passed through a further likelihood test between the original classification for a given segment and the Silence class, the result of which determines whether they are "Near" (high probability of being that class) or "Far" (low probability, they may be silence instead). Given the large number of acronyms and labels of various kinds, we provide a listing of relevant LENA® abbreviations in Table 1.

After this broad speaker classification step, Female or Male Adult "Near" segments (FAN and MAN) are further processed using an adaptation of the Sphinx Phone Decoder (Lamere et al., 2003) in order to form an automated estimate of the number of words in each segment (Adult Word Count, or AWC). Key Child (CHN) segments are further processed to sub-classify regions in them into vegetative noises, crying, and speech-like vocalizations. LENA® provides counts (child vocalization count, or CVC) and durations for this last speech-like sub-segment category. A further metric, Conversational Turn Counts (CTC), reflects the number of alternations between an adult and the key child (or vice versa), bounded by a maximum 5s of non-speech.

Previous validation work. A recent systematic review (A. Cristia et al., under review) found 23 papers containing 28 studies that reported on the accuracy of the LENA® system's labels and/or derived metrics (AWC, CVC, CTC). They conclude that there are:

> "reasonably good results [overall]: over 61% for recall and precision based on 11-12 non-independent studies; correlations for AWC mean r=.79, on n=11, with

Table 1

A partial listing of common LENA abbreviations and their meanings.

Abbreviations	Meanings
FAN, MAN, CHN, CXN	Basic "meaningful speech" (near and clear speech) categories used by LENA for further processing: Female Adult Near, Male Adult Near, Key Child Near and Other Child Near categories respectively.
NON, TVN, OLN, SIL	Basic non-speech categories: Noise Near, Television Near, Overlap Near, Silence.
FAF, MAF, etc.	"Far" (low probability) versions of each Near category.
Key child	Child wearing recorder
AWC	Adult Word Count (estimated within FAN and MAN vocalizations)
CVC	Child Vocalization Count (estimated for non-cry, non-vegetative portions of CHN)
CTC	Conversational Turn Count (estimated for turns between FAN or MAN and CHN)

a mean RER [i.e., Relative Error Rate]=10% on n=11; CVC mean r=.76, n=5, with a mean RER=1% on n=5. The exception to this general trend towards good performance was CTC, with a mean r=.31, n=5, RER=-64% on n=2."

The systematic review also identified several limitations of previous validation work. First, for the majority of included studies, the validitation component was not fully evaluated by peer review. Even if the study may have appeared in a pee-reviewed journal, the validation in itself was often a secondary goal to support a different research objective, and therefore often lacked methodological details or even full results. For instance, Seidl et

al. (2018) report on validation of LENA® labels among children at familial risk for autism in a one-paragraph appendix to the paper, which only mentions confusions between female adult and child. This leaves unclear whether confusions between key child and any other category (other child, male adult, silence, etc.) were ignored considered to be as errors. While this approach may be reasonable for a given study's research goals, it has the undesirable side effect of creating the impression that LENA® metrics are widely validated, while in fact validation methods may not have been reported or evaluated in detail. Second, previous studies typically did not take silence, noise, or overlap into account in the reported confusion matrices or other accuracy measures, particularly within segments. That is, if a LENA® segment labeled "key child" contained one second of silence and two seconds of speech by the key child, the full three second clip may be tagged as "correct" though it was only 67% correct, leading to an overestimation of the accuracy of the "key child" label.

Third, a majority of previous validation studies used the LENA® output itself to select the sections that would be annotated for validation (in A. Cristia et al., under review, this held for 14/25 studies that specified the method of selection). For instance, clips may have been selected for manual annotation on the basis of high AWC and/or CTC according to the algorithm. This unfortunately leads to biased sampling: Since LENA® only counts words within FAN and MAN segments and conversational turns involving FAN/MAN alternations with CHN in close proximity, high AWC or CTC can only occur in sections of the recording that are "clean" enough for the algorithm to parse; otherwise, most of the section would have been classified as overlap (OLN), which does not count towards AWC or CTC. This would tend to bias these reports toward a higher level of accuracy than would be obtained across the full recording.

Fourth, previous validation work has typically focused on a single corpus, participant population, age range, and language. As a result, although considerable variation in performance has sometimes been reported (Canault et al., 2016; e.g., Gilkerson et al., 2016) it is difficult to assess whether a numerical difference in accuracy found is significant, and if

so, whether this is due to a difference in the way the corpus was constituted and annotated, rather than on how LENA® fares with that population, age range, and language.

The present work. We sought to assess the validity of the output provided by LENA® through an approach that complements the preceding literature. Specifically, we report an evaluation of all speech labels and the silence (SIL), overlap near (OLN), and overlap far (OLF) non-speech labels; as well as the system's key derived metrics: adult word count (AWC), conversational turn count (CTC), and child vocalization count (CVC). We aim to address several of the limitations found in the body of previous work.

First, to maximally avoid potential bias in our annotations, we used random or periodic sampling (detailed below) to choose which sections of daylong recordings to annotate, and did not give annotators access to the LENA® output. Second, to allow assessment of the accuracy of the segmentation itself as well as categorical labeling, evaluation is done at the level of 10 ms frames rather than the LENA® segments. This allows us to capture a much finer-grained representation of the auditory environment (i.e., if LENA® classified a 2 s audio segment as FAN, but .8 s of this was actually non-speech noise or a different talker, in our analysis LENA® would be credited only for the proportion that was correct).

Third, to gain traction on generalizability, rather than focusing on a single sample that either mirrors or diverges from LENA®s original population, we included five corpora. Three corpora sampled from the same population, language, dialect, and age group the LENA® software was developed with. A fourth corpus was chosen to allow an extension to a different dialect of English. The fifth corpus constituted an extension to a totally different recording condition (a rural setting, with large families and many children present, in a typologically different language). The age range also varies a great deal, and it is slightly higher in this last corpus.

Finally, the present study relies on a collaborative effort across several labs. The annotation pipeline was identical for four of the corpora, and conceptually comparable to the fifth (as detailed below). This allows us to more readily answer questions regarding

differences in reliability as a function of e.g. child age and language. This approach also let's us better infer the likelihood with which our results will generalize to other corpora, provided the annotation scheme is conceptually comparable.

Methods

Corpora. The data for the evaluation comes from five different corpora, annotated in the context of two research projects. The largest one is the ACLEW project (E. Bergelson et al., 2017; Soderstrom et al., in progress); in this paper we focus on four different corpora of child daylong recordings that have been pooled together, sampled, and annotated in a coordinated manner. These four corpora are: the Bergelson corpus ("BER") from US English families from the upstate New York area (E. Bergelson, 2016), the LuCiD Language 0–5 corpus ("L05") consisting of English-speaking families from Northwest England (C. F. Rowland, Bidgood, Durrant, Peter, & Pine, 2018), the McDivitt and Winnipeg corpora ("SOD") of Canadian English families (McDivitt & Soderstrom, 2016), and the Warlaumont corpus ("WAR") of US English from Merced, California (A. Warlaumont, Pretzer, Walle, Mendoza, & Lopez, 2016). Some recordings in BER, and all recordings in SOD and WAR are available from HomeBank repository (VanDam et al., 2016). The second project contains a single corpus collected from Tsimane' speaking families in Bolivia ("TSI"; C. Scaff, Stieglitz, Casillas, & Cristia, in progress). Socioeconomic status varies both within and across corpora. Key properties of the five corpora are summarized in Table 2.

Despite these differences, all five corpora consists of long (4–16 hour) recordings collected as children wear a LENA® recorder in a LENA® vest throughout a normal day and/or night. For the four ACLEW corpora, out of the 106 recorded participants, daylong recordings from 10 infants from each corpus were chosen for manual annotation, selected to represent a diversity of ages (0–36 months) and socio-economic contexts. In the SOD corpus, sensitive information was found in one of the files, and thus one child needed to be excluded. The tenth day for this corpus was a second day by one of the 9 included children. From each

Table 2

Key properties of the five corpora

Corpus	Children	Clips	Clip duration (seconds)	Mean Age [range] (months)	Location
WAR	10	150	120	6.3 [3-9]	Western US
BER	10	150	120	11.2 [7-17]	Northeast US
SOD	9	150	120	12.3 [2-32]	Western Canada
L05	10	150	120	20 [11-31]	Northwest England
TSI	10	272	60	34 [15-58]	Northern Bolivia

daylong file, fifteen 2-minute non-overlapping audio (with a 5-minute context window) were randomly sampled from the entire daylong timeline for manual annotation. In total, this lead to 20 hours of audio, and 4.6 hours of annotated speech/vocalizations (collapsing across all speaker categories).

The TSI corpus consisted of 1 or 2 recordings from 12 children, out of the 25 children recorded from field work that year; the other 13 had been recorded using other devices (not the LENA® hardware). From these files, 1-minute segments were sampled in a periodic fashion. That is, for each recording, we skipped the first 33 minutes to allow the family to acclimate to the recorder, and then extracted 1 minute of audio (with a 5-minute context window) every 60 minutes, until the end of the recording was reached. This resulted in a total of 4.5 hours of audio, and 0.7 hours of speech (collapsing across all speaker categories).

We chose to sample 1 or 2 minutes at a time (TSI, and ACLEW corpora, respectively) because conversations are likely to be bursty (Goh & Barabási, 2008). That is, it is likely the case that speech is not produced at a periodic rate (e.g., one phrase every 20 seconds), but rather it occurs in bursts (a conversation is followed by a long period of silence between the conversational partners, followed by another bout of conversation, perhaps with different interlocutors, followed by silence, and so on). In this context, imagine that you sample a

5-second stretch. If you find speech in that stretch, then it is likely you have by chance fallen on a conversation bout; if you do not find speech, then you have likely found a silence bout. If you were to extend that selection out to several minutes, then it is likely that you will simply add more material from the same type (i.e. conversation bout or silence bout). As a result, any sampling method that favors medium-sized stretches (5-15 minutes) will tend to end up with samples that are internally homogeneous (throughout the 5 minutes, there is a conversation, or there is silence throughout). If smaller clips are sampled out, this heterogeneity is still captured, but (keeping the total length of audio extracted fixed) the number of clips that can be extracted is larger, thus likely increasing the likelihood that results will generalize to a new section of the audio.

In the 5 corpora, the 1- or 2-min samples were annotated for all hearable utterance boundaries and talker ID. In ACLEW corpora (Bergelson et al., 2019 for the general annotation protocol; cf. Casillas et al., 2017; Soderstrom et al., in progress, for an introduction to the databases), talker IDs reflected unique individual talkers, but were coded in such a way to readily allow mapping onto LENA®s talker categories (e.g. key child, other child 1, female adult 1, female adult 2). In the TSI corpus, only the key child and one female adult whose voice recurred throughout the day were individually identified, with all other talkers being classified on the basis of broad age and sex into male adult, female adult, and other children. The ACLEW datasets had other coding levels which will not be discussed here.

Processing. Several different time units are needed to clarify how each metric is calculated (see Figure 1). Clips refer to the 1- or 2-minute samples extracted from recordings (TSI corpus and ACLEW corpora, respectively). This is the basic unit at which child vocalization counts and conversational turn counts can be established. In addition, since most previous work evaluating adult word counts did so at the clip level, we do so here as well.

The other metrics require a more detailed explanation, conveyed graphically in Figure

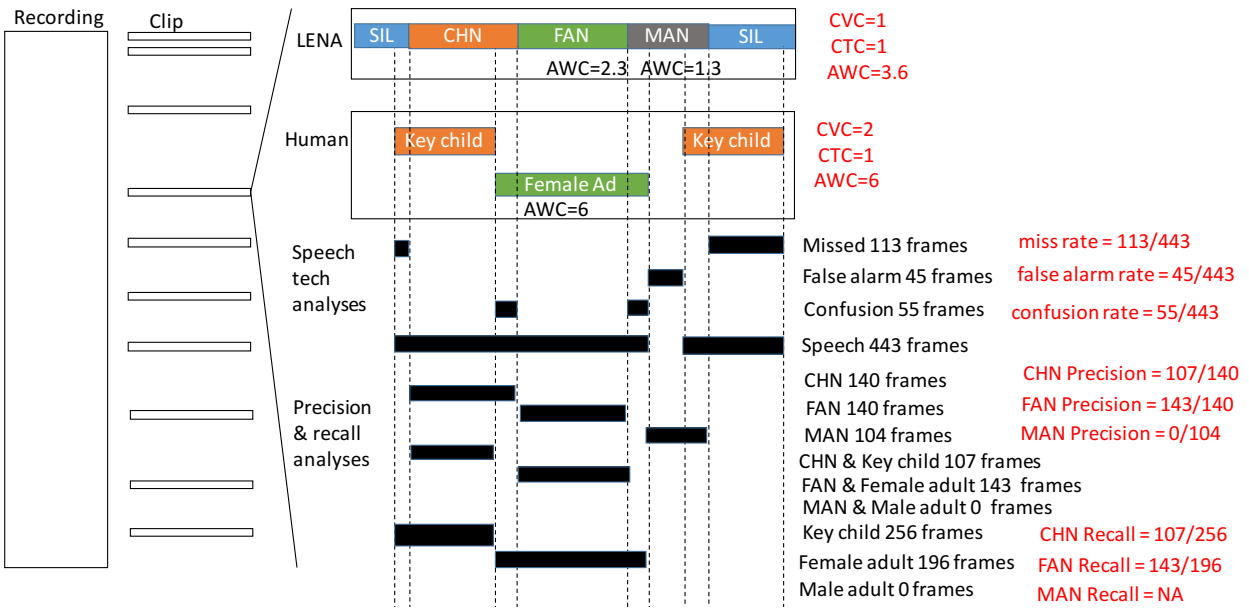

Figure 1. Levels at which performance is evaluated. Adult Word Count (AWC), Child Vocalization Count (CVC), and Child Turntaking Count (CTC) are calculated at the level of the 1- or 2-minute long audio extracts (clips). Misses, false alarms, confusions as well as class precision and recall depend on 10-ms frames, and are totalled both at the level of individual clips and over the full audio.

1. The stretch of time that has been assigned to a speech or non-speech class by LENA® is a *segment*. In one clip, there may be just one long segment (e.g., the whole clip has been assigned to Silence by LENA®); or there may be more (e.g., the first 5 seconds are attributed to the key child, then there is a 50-second Silence segment, and the final 5 seconds are attributed to a Female Adult). In the system's automated analysis, only one of these categories may be active at a given point in time. In contrast, colloquially, utterances or vocalizations to refer to stretches of speech detected by humans and assigned to different talkers. To be clear: clips may have zero or more utterances. Unlike LENA®, however, a given point in time may be associated with multiple speakers.

Given that there need not be a one-to-one correspondence between LENA® segments and human utterances, we need to define smaller time units that can be used to check for

classification agreement. In this paper, we use 10 ms *frames*. This is the basic time unit used for all classification accuracy estimations, which are introduced in more detail in the next subsection.

LENA® classification accuracy. Our first goal was to establish LENA® talker tag accuracy, particularly for the four broad LENA® talker categories (key child, other child, female adult, male adult; or CHN, CXN, FAN, MAN), but taking into account other categories (with some limitation on their interpretation). We calculated accuracy in two complementary ways. First, we used three frame-based standard metrics of speech and talker segmentation to allow direct comparison with other systems in the speech technology literature (False Alarm Rate, Miss Rate, Confusion Rate). We also use Diarization Error Rate, which is derived by summing the first three metrics; together these provide a stringent and standard test of accuracy. Second, we used frame-based precision and recall of each category to provide an intuitive representation of the error patterns shown by this system.

Speech and talker segmentation metrics. The original coding was converted using custom-written python scripts into a standard adaptation of the "Rich Transcription Time Mark" (rttm) format (Ryant et al., 2019), which indicates, for each vocalization or segment, its start time, duration, and speaker. This representation was used in pyannote.metrics (Bredin, 2017) to compute four standard diarization metrics: rate of false alarm for speech, rate of misses for speech, rate of confusion between talkers, and the derived diarization error rate (DER). These are calculated with the following formulas at the level of each clip, where FA (false alarm) is the number of frames during which there is no talk according to the human annotator but during which LENA® found some talk; M (miss) is the number of frames during which there is talk according to the human annotator but during which LENA® found no talk; C (confusion) is the number of frames correctly classified as LENA® as containing talk, but whose voice type has not been correctly identified (when the LENA® model recognizes female adult speech where there is male adult speech for instance), and T is the total number of frames that contain talk according to the human annotation:

Table 3

*Correspondances between LENA and our human annotation tags for each talker type. Additional analyses remove one or both of the last two rows. *Electronic voices were only annotated in the ACLEW dataset. N.B. Although some Tsimane' families listen to the radio, radio speech was not annotated in the TSI corpus.*

Talker	LENA	Human
Key Child	CHN	CHI
Other Child	CXN	OCH
Female Adult	FAN	FA
Male Adult	MAN	MA
Electronics	TVN*	E*
Overlap	OLN	OL

- **F**alse **A**larm rate = FA/T (T=**T**otal # of frames that contain talk),
- **M**iss rate = M/T,
- **C**onfusion rate = C/T,
- **D**iarization **E**rror **R**ate (DER) = (FA+M+C)/T

In the human annotation, there is no class representing overlapping speech as such. For the sake of completeness and full comparison with the LENA® model, if two or more different speech sources were active at the same time according to the human annotators, these frames have been mapped to the class "overlap" post hoc. This allows us to compare this Overlap class to the system's OLN (and, for the precision/recall analysis introduced next, OLF) by the LENA® model. Therefore, in the most complete analysis, the confusion rate is computed based on the human-LENA® matches in Table 3.

However, our overlap category is not defined identically to the LENA® overlap category. For LENA®, overlap between any two categories is labeled OLN – i.e., Noise + TV would be

counted towards overlap as would FAN+FAN; whereas for us, only overlap between two talker categories (e.g., key child and female adult, noise was not coded) counts as overlap. Similarly, the TVN LENA® class is not equivalent to the electronic speech tag in the ACLEW coding, because the former also includes music, singing, crowd noise and any other sound coming from a TV or another electronic source, whereas the latter only includes speech from an electronic source. Therefore, additional analyses map these classes onto "silence" post hoc, so as to not penalize confusions involving them.

Precision and recall. This evaluation looks in more detail at the pattern of errors, by assessing how LENA® and human annotators agreed and disagreed. In both precision and recall, the numerator is the intersection between a LENA® tag and a human tag (e.g., the number of frames that LENA® classified as CHN and the annotator classified as Key child). The denominator differs: To calculate precision, we divide that number by the total number of frames attributed to a category by LENA®, whereas for recall, we divide by the total number of frames attributed to a category by the human annotator.

CVC and CTC evaluation. From the human annotation, each vocalization by the key child counted towards the total Child Vocalization Count (CVC) for a given clip. For the Conversational Turn Count (CTC), a sequence of key child and any adult (or vice versa) within 5 seconds counted towards the clip total CTC. The Pearson correlation across LENA® and human estimations was calculated.

Users may also wish to interpret the absolute number of vocalizations or turns found by LENA®. Therefore, it is important to also bear in mind absolute error rates, relative error rates, and relative absolute error rates. Despite the similarity in their names, these three metrics provide different information. The *absolute error rate* tells us, given a LENA® estimate, how close the actual number may be. It is calculated as NL-NH, where NL is the number according to LENA® and NH is the number according to humans. By averaging across clips, we then get an idea of the bias towards overestimation (if this number is positive) or underestimation (if this difference is negative).

The *relative error rate* puts this bias in relation to the actual number of vocalizations tagged by the human coder: (NL-NH)/NL. For instance, imagine that we find that LENA® errs by 10 vocalizations on average according to the average absolute error rate; this means that, on average across short clips like the ones used here, the numbers by LENA® would be off by 10 vocalizations. By using the relative error rate, we can check whether this seemingly small difference is indeed small relative to the actual number found. I.e, an error of 10 vocalizations would be less problematic if there were 100 vocalizations on average (LENA® would be just 10% off) than if there were 10 (LENA® would be doubling the number of vocalizations). As with the absolute error rate, the sign of this difference indicates whether LENA® tends to over- or under-estimate these counts.

Finally, the *relative absolute error rate* is calculated with the formula abs(NL-NH)/NL, where abs indicates that one takes the absolute of the difference. As a result, it cannot be used to assess systematic under- or over-estimation biases, but rather gives an idea of how accurate the estimates are at the clip level (statistically speaking). To convey this intuitively, one could find absolute error rates and relative error rates that are 0 because half of the samples are -100 vocalizations off (for the absolute error rate) or -100% off (for the relative error rates), with the other half behave in the exact opposite fashion.

AWC evaluation. For the AWC portion of this evaluation, we could only use transcriptions from the four ACLEW corpora, since the TSI corpus has not been transcribed (and thus lacks word counts). Annotators for the four ACLEW corpora were proficient in the language spoken in the daylong recording, and transcribed all adult speech based using canonical lexical forms (e.g. “wanna”, not “want to”) in keeping with minCHAT format (MacWhinney, 2017).

Reference adult word counts were determined by counting all unambiguously transcribed words spoken by adult talkers. This was achieved by first discarding all non-lexical transcript entries such as non-linguistic communicative sounds, paralinguistic markers, and markers indicating incomprehensible speech. In addition, all utterances from

the key child and other children were omitted from the Adult Word Count(AWC). The remaining orthographic entries separated by whitespaces were then counted as gold standard target words for LENA® to detect.[1]

For the LENA® Adult Word Counts, the regions sampled for manual annotation were not guaranteed to perfectly align with LENA® segments. Of all LENA® segments found within the extracted clips, 14% straddled a clip boundary (i.e., the segment began before the clip started; or it ended after the extracted clip ended). To match LENA® AWCs with the annotated word counts, words from these straddling LENA® segments were included in proportion to the length of the LENA® segment included in the clip (e.g., if 10% of the duration of a LENA® segment fell within a clip, 10% of the LENA® AWC estimate for that segment was included in the LENA® word count estimate for that clip). AWC was evaluated using Pearson correlations and error rates, similarly to CVC and CTC.

Results

Before starting, we provide some general observations based on the manual human annotations. Silence was extremely common, constituting 71% of the 10 ms frames. In fact, 34% of the 1-2 minute clips contained no speech by any of the speaker types (according to the human annotators). As for speakers, female adults made up 16% of the frames, the child contributed to 6%, and male adult voices, other child voices, and electronic voices were only found in 1% of the frames each. Overlap made up the remaining 3% of frames. The following consequences ensue. If frame-based accuracy is sought, a system that classifies every frame as silence would be 71% correct. This is of course not what we want, but it indicates that systems well adapted to this kind of speech should tend to have low false alarm rates, being very conservative as to when there is speech. If the system does say there is speech, then it had better say that this speech comes from female adults, who provide a great majority of

[1]It turned out that one child in the (otherwise English) SOD corpus was learning French. Given our definition of orthographic words which is not language-specific processing, we have included this child to increase power, but results without them are nearly identical.

the speech, nearly 3 times as much as the key child and 10 times more than other children or male adults. In fact, given that speech by male adults and other children is so rare, a system that makes a lot of mistakes in these categories may still have a good global performance, because males and other children jointly accounted for only 2% of the frames.

LENA® classification accuracy: False alarms, misses, confusion. Our first analysis is based on standard speech technology metrics, which put errors in the perspective of how much speech there is. That is, if 10 frames are wrong in a file where there are 100 frames with speech, this is a much smaller problem than if 10 frames are wrong in a file where there is 1 frame with speech. In other words, these metrics should be considered *relative error* metrics.

However, one problem emerges when there is no speech whatsoever in a given file. In the speech technology literature, this is never discussed, because most researchers working on this are basing their analyses on files that have been selected to contain speech (e.g., recorded in a meeting, or during a phone conversation). We still wanted to take into account clips with no speech inside because this was central for our research goals: We need systems that can deal well with long stretches of silence, because we want to measure in an unbiased manner how much speech (and silence!) children hear. Unfortunately, in the 34% of clips that had no speech whatsoever, the false alarm, miss, and confusion rates are all undefined, because the denominator is zero. To be able to take clips with no speech into account, we defined the following rules, to deal with the cases that arose. First, if a clip had no speech according to the human annotator, while LENA said there was speech, then the false alarm rate was 100%, and the miss and confusion rates were zero. Second, if on the contrary, both the human annotator and LENA said there was no speech, then all the error rates were zero. Finally, in some cases there was just a little speech; in this case, the denominator was very small, and therefore the ratio for these two metrics ended up being a very large number. To be maximally informative, we report results in three ways: (1) Overall false alarm, miss, and confusion rates over all clips together, thus giving more weight to clips with more speech; (2)

As means across clips, which represent central tendency when giving equal weight to clips with more versus less or no speech; and (3) since means are not robust to outliers, we also report the median across all clips.

As mentioned briefly above, there were, *a priori*, several ways of analyzing the data:

- collapsing LENA® "near" and "far" together (e.g., mapping CHN and CHF onto a single CH category);
- distinguishing the "near" and "far" categories (e.g., treating CHN and CHF as distinct "speakers");
- excluding TV as a speaker category, since it is conceptually not identical to the electronic voices detected by ACLEW human annotators; in this case, the TV labels would be mapped post hoc to "silence", as would the electronic voices in the human annotation;
- omitting OLN as a speaker category, since it is not conceptually identical to the overlap derived from humans' annotating different speaker categories; in this case, the LENA® system's OL labels would be mapped post hoc to "silence", as would the regions of overlap in the human annotation.

We thought it most informative to report on several of these options, albeit briefly. We start with the situation that yields the best LENA® performance: mapping electronic voices and overlap in the human annotation onto silence, so that the categories found in the human annotation are FEM, MAL, CHI, OCH, alongside using only CHN, FAN, MAN, and CXN as speakers in the LENA® annotation, (with all "far" categories, TVN, and OLN all mapped onto silence; see Tables 1 and 3). Calculated in this way, the LENA® system's *false alarm* rate (i.e., tagging a non-silence category when there wasn't one) was 13% overall, with a mean across all clips of 26% and a median of 5%. The *miss rate* was 56% overall, with a mean across all clips of 37% and a median of 37%. The *confusion rate*, as mentioned above, is only calculated for the correctly detected speech (i.e., not speech that was missed, which counts towards the miss rate, nor the speech that was falsely identified, which is part of false

alarms). The confusion rate was very low, 9% overall, with a mean across all clips of 8% and a median of 6%. These three metrics (*false alarms*, *misses*, and *confusions*) are summed to yield the *diarization error rate*. The overall diarization error rate was 77%, with a mean across all clips of 68% and a median of 69%.

For simplicity's sake, for the subsequent re-analyses using different mappings from talkers to categories as bulleted above, we focus on only *medians* over all clips, capturing performance across all samples without being overly sensitive to outliers. In the second-best performing case, electronic voices in the human annotation are still mapped onto silence but overlap is not, so that the human categories considered were CHI, FEM, MAL, OCH, and overlap; and the LENA® categories considered were CHN, FAN, MAN, CXN, and OLN (with all "far" classes and TVN mapped onto silence). In this setting, the LENA® system's false alarm, miss, and confusion rate medians were 32%, 22%, and 27% respectively, for a total median diarization error rate of 81%. Although the miss rate is reduced when OLN is considered speech, overall performance degraded because OLN was not picking out the same regions as the overlapping speech found in the human annotations, leading to considerably higher false alarm and confusion rates.

Next, we allowed the electronic voices segmented by human annotaters, and the LENA® system's TVN speaker categories to be included in the evaluation (rather than mapping them all to silence), such that the human categories considered were CHI, FEM, MAL, OCH, overlap, and electronic; and the LENA® categories considered were CHN, FAN, MAN, CXN, OLN, and TVN. the LENA® system's false alarm, miss, and confusion rate medians were 40%, 19%, and 29% respectively, for a total median diarization error rate of 88%. Again, the miss rate drops slightly as we allow one more class to be considered speech, but this does not make up for the higher false alarm and confusion rate, likely because TVN is not defined similarly to the electronic speech segmented by ACLEW annotators.

Finally, we included the maximum possible number of categories. The human categories considered were still CHI, FEM, MAL, OCH, overlap, and electronic, alongside

the fuller range of LENA® categories: CHN, FAN, MAN, CXN, OLN, TVN, CHF, FAF, MAF, CXF, OLF, and TVF. the LENA® system's false alarm, miss, and confusion rate medians were 71%, 7%, and 41% respectively, for a total median diarization error rate of 119%. Thus, performance degrades considerable when all the "far" classes are treated as speech, leading to huge false alarm rates and increased confusion rates.

LENA® classification accuracy: Precision and recall. By now, we have established that the best performance (when "far" labels such as CHF and OLF are mapped onto silence, as are TVN and OLN), the median diarization error rate is about 69%, due mainly to missing speech (37%), with false alarms (5%) and confusion between talker categories (6%) constituting a relatively small proportion of errors. These metrics may be insufficient for our readers for two reasons. First, these metrics give more importance to correctly classifying segments as speech versus non-speech (false alarms + misses) than confusing talkers (confusion). Second, many LENA® adopters use the system not to make decisions on the sections labeled as non-speech, but rather on sections labeled as speech, and particularly those labeled adults and key child. The metrics above do not give more importance to adults and the key child, and they do not give us insight on the patterns of error made by the system.

We therefore turn to precision and recall. Looking at precision of speech categories is crucial for users who interpret the LENA® system's estimated quantity of adult speech or key child speech, as low precision means that some of what LENA® called e.g. key child was not in fact the key child, and thus it is providing overestimates. Looking at recall may be most interesting for adopters who intend to employ LENA® as a first-pass annotation: the lower the recall, the more is missed by the system and thus cannot be retrieved (because the system labeled it as something else, which will not be inspected given the original filter).

This subsection shows confusion matrices, containing information on precision and recall, for each key category. For this analysis, we collapsed over all human annotations that contained overlap between two speakers into a category called "overlap". Please remember

that this category is not defined the same way as the LENA® overlap category. For LENA®, overlap between any two categories falls within overlap – i.e., CHN+TV would be counted towards overlap; whereas for us, only overlap between two talker categories (e.g., key child and female adult) counts as overlap.

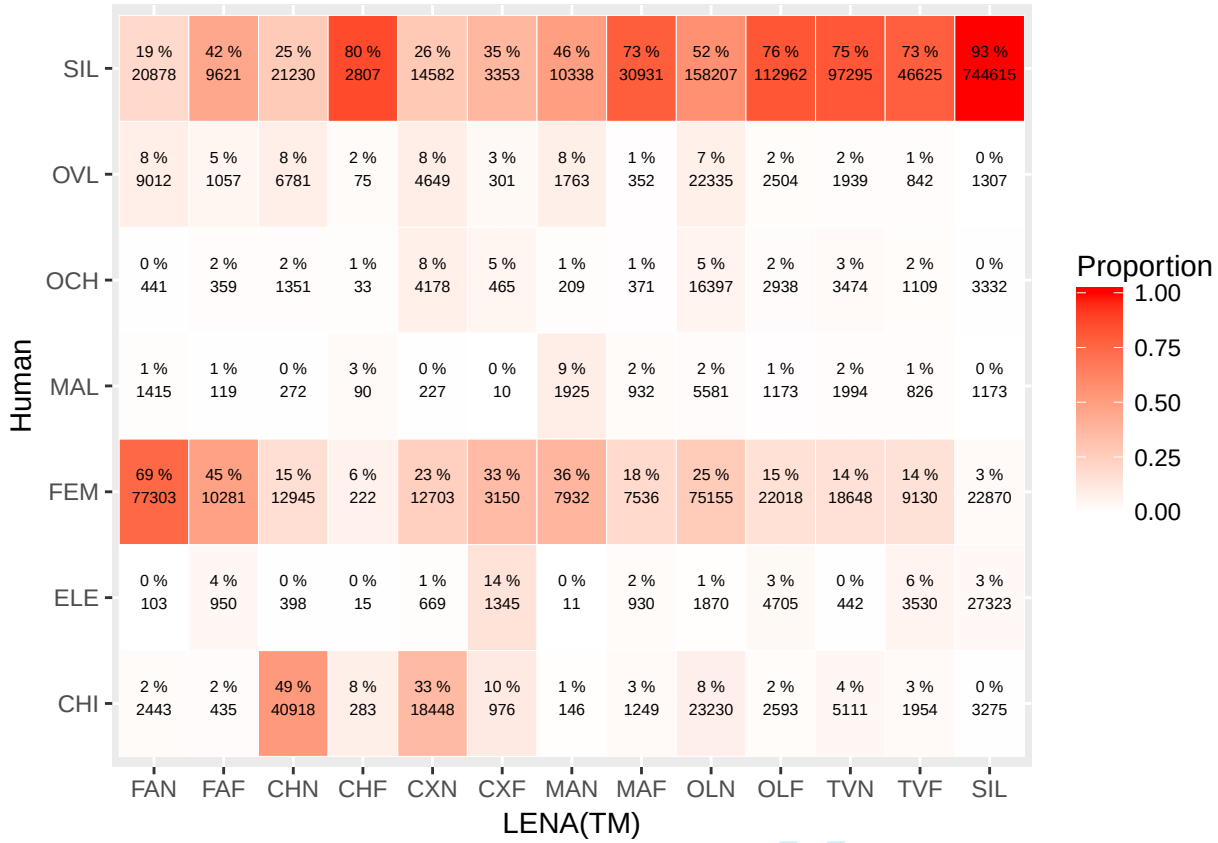

Figure 2. Precision: Confusion matrix between LENA (x axis) and human annotations (y axis). In each cell, the top number indicates the percentage of all frames in that LENA category (column) that are labeled as a given class by the human (row); cells in a given column add up to 100%. The number below indicates number of frames in that intersection of LENA and human classes.

LENA® classification accuracy: Precision. We start by explaining how to interpret one cell in Figure 2: the darkest cell in column 1. This is the cross of the human category (i.e. row) FEM and the LENA® category (i.e. column) FAN; when LENA® tagged a given frame as FAN, this corresponded to a frame tagged as being a female adult by the

human 69% of the time. This category, as mentioned above, was the most common speaker category in the audio, so that over 77k frames (representing 69% of the frames tagged as FAN by LENA®) were tagged as a female adult by both the human and LENA®. The remaining 31% of frames that LENA® tagged as FAN were actually other categories according to our human coders: 19% were silence, 8% were in regions of overlap between speakers or between a speaker and an electronic voice, and 4% were confusions with other speaker tags. Inspection of the rest of the confusion matrix shows that, other than silence, this is the LENA® tag with the greatest precision.

Precision for CHN comes in second place, at 49%; thus, nearly half of the frames labeled as the key child are, in fact, the key child. The majority of the frames that LENA® incorrectly tagged as being the key child are actually silence (or rather, lack of speech) according to the human annotator (25%), with the remaining errors being due to confusion with other categories. ~15% of them are actually a female adult; 2% are another child, and 8% are regions of overlap across speakers, according to our human coders.

MAN and CXN score similarly, 9% and 8% respectively, and poorly. That is, less than 10% of what LENA® tagged as these speakers actually correspond to them. As with the key child, most errors are due to LENA® tagging silent frames as these categories. However, in this case confusion with other speaker tags was far from negligible. In fact, the most common speaker tag in the human annotation among the regions that LENA® tagged as being MAN was actually female adult speech (36%). It was also not uncommon to find a CXN tag for a frame human listeners identified as a female adult (23%) or the key child (8%). In a nutshell, this suggests extreme caution before undertaking any analyses that rely on the precision of MAN and CXN, since most of what is being tagged with these talker codes by LENA® is silence or other speakers.

Another observation is that the "far" tags of the speaker categories do tend to more frequently correspond to what humans tagged as silence (63%) than the "near" tags (41%), and thus it is reasonable to exclude them from consideration. The relatively high proportion

of near LENA® tags that correspond to regions that humans labeled as silence could be partially due to the fact that the LENA® system, in order to process a daylong recording quickly, does not make judgments on short frames independently, but rather imposes a minimum duration for all speaker categories, padding with silence in order to achieve it. Thus, any key child utterance that is shorter than .6 s will contain as much silence as needed to achieve this minimum (and more for the other talker categories). Our system of annotation, whereby human annotators had no access whatsoever to the LENA® tags, puts us in an ideal situation to assess the impact of this design decision. That is, any manual annotation that starts from the LENA® segmentation would bias the human annotator to ignore such short interstitial silences to a greater extent than if they have no access to the LENA® tags.

These precision analyses shed light on the extent to which we can trust the LENA® tags to contain what the speaker tag name indicates. We now move on to *recall*, which indicates a complementary perspective: how much of the original annotations attributed to a given class was captured by the corresponding LENA® class.

LENA® classification accuracy: Recall. Again, we start with an example to facilitate the interpretation of Figure 3. As seen at the intersectin of human CHI (row 1) and LENA® CHN (column 3), the best performance for a talker category for recall is CHN: 40% of the frames humans tagged as being uttered by the key child were captured by the LENA® under the CHN tag. Among the remainder of what humans labeled as the key child, 18% was captured by the LENA® system's CXN category and 23% by its OLN tag, with the rest spread across several categories.

This result suggests that an analysis pipeline that uses the LENA® system to capture the key child's vocalizations by extracting only CHN regions will get nearly half of the key child's speech. If additional manual human vetting is occuring in the pipeline, researchers may find it fruitful to include segments labeled as CXN, since this category actually contains a further 18% of the key child's speech. Moreover, as we saw above, over 33% of these

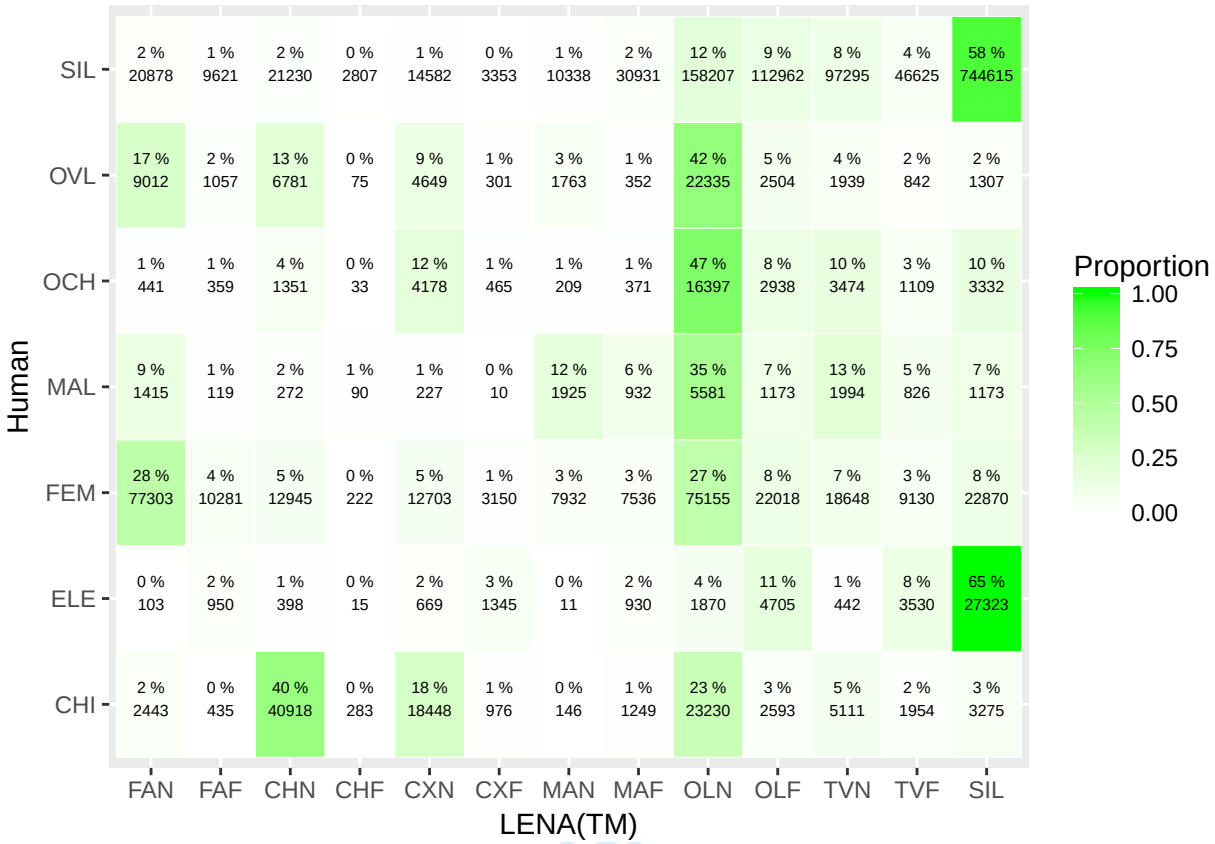

Figure 3. Recall: Confusion matrix between LENA (x axis) and human annotations (y axis). In each cell, the top number indicates the percentage of all frames that a human labeled as a given class (row) which were recovered in a given LENA category (column); cells in a given row add up to 100%. The number below indicates number of frames in that intersection of LENA and human classes.

LENA® tags corresponds to the key child, which means that human coders re-coding these regions could filter out the 67% that do not.

Many researchers also use the LENA® as a first pass to capture female adult speech through the FAN label. Only 28% of the female adult speech can be captured this way. Unlike the case of the key child, missed female speech is classified into many of the other categories, and thus there may not exist an easy solution (i.e., one would have to pull out all examples of many other categories to get at least half of the original female adult). However, if the goal is to capture as much of the female speech as possible, a reasonable solution

would be to include OLN regions, since these capture a further 27% of the original female adult speech and, out of the OLN tags, 25% are indeed female adults (meaning that human annotators re-coding these regions need to filter out 3/4 clips, on average).

For the remaining two "near" speaker labels (MAN, CXN), recall averaged 12%, meaning that less than a quarter of male adult and other child speech is being captured by LENA®. In fact, most of these speakers' contributions are being tagged by the LENA® as OLN (mean across MAN and CXN 35%) or TV (mean across MAN and CXN is 13%), although the remaining sizable proportion of misses is actually distributed across many categories.

Finally, as with precision, the "far" categories show worse performance than the "near" ones. It was always the case that a higher percentage of frames is captured by the near rather than the far labels. For instance, out of all frames attributed to the key child by the human annotator, 40% were picked up by the LENA® CHN label versus 0% by the LENA® CHF label. This result provides further support that when sampling LENA® daylong files using the LENA® software, users likely need not take the "far" categories into account.

Child Vocalization Counts (CVC) accuracy. To calculate Child Vocalizaton Counts, given the inaccuracy of "far" LENA® tags and our goal of following the LENA® system procedure, we only counted vocalizations attributed to CHN and ignored those attributed to CHF. As shown in Figure 4, there is a strong association between clip-level counts estimated via the LENA® system and those found in the human annotations: the Pearson correlation between the two was r = 0.80 (p = 0.00) when all clips were taken into account, and r = 0.75 (p = 0.00) when only clips with some child speech (i.e., excluding clips with 0 counts in both LENA® and human annotations) were considered. This suggests that the LENA® system captures differences in terms of number of child vocalizations across clips well.

The *absolute error rate* ranged from -47% to 18%, with a mean of -2.92% and a median of 0%. Since these numbers can be affected by silent clips, in which both humans and LENA®

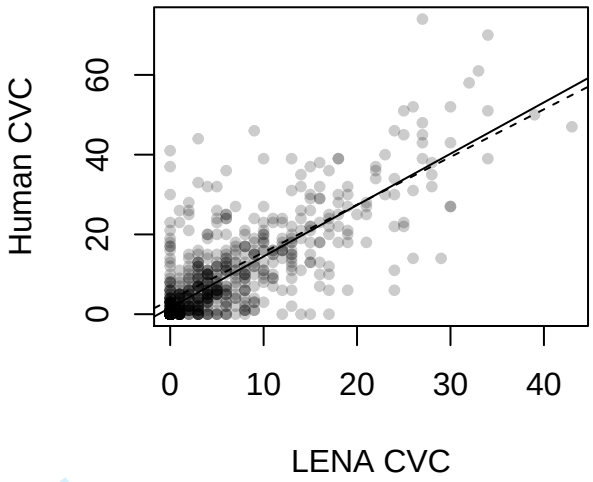

Figure 4. Child Vocalization Counts according to LENA (x axis) and humans (y axis). Each point represents the CVC totaled within a clip. The solid line corresponds to a linear regression fit to data from all clips; the dashed line corresponds to an analysis excluding clips where both the human and LENA® said zero child vocalizations.

may trivially agree on there not being any vocalizations, we also calculated the absolute error rate excluding these clips. In this case, the mean was -5.31% and the median was -4%.

Next, we considered *relative error rates*, which require the number in the denominator to be non-null. For this analysis, therefore, we need to remove the 379 clips in which the human annotator said there were no child vocalizations whatsoever. When we do this, the mean relative error rate ranged from -100% to 800%, with a mean of -19.66% and a median of -43.75%. Finally, the *absolute relative error rate* ranged from 0% to 800%, with a mean of 66.84% and a median of 50%.

Together, these data suggest that LENA tends to underestimate vocalization counts, particularly when only clips with some speech are considered. This understimation is quite systematic and it appears to be around half of the actual counts.

Conversational Turn Counts (CTC) accuracy. Again, we only considered "near" speaker categories in the turn count, and applied the same rule the LENA® does, where a turn can be from the key child to an adult or vice versa, and should happen within 5

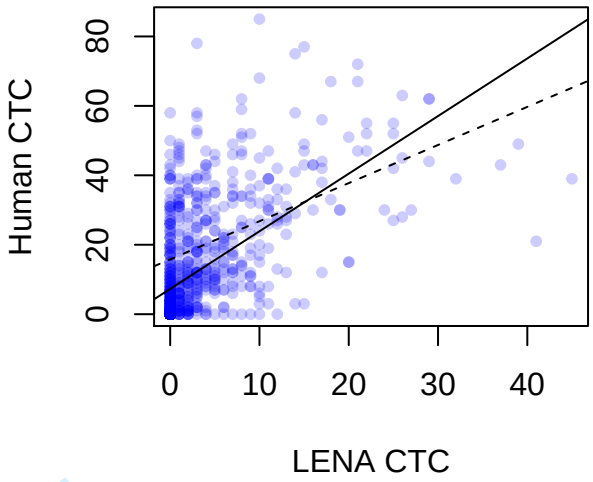

Figure 5. Conversational Turn Counts according to LENA (x axis) and humans (y axis). Each point represents the CTC totaled within a clip. The solid line corresponds to a linear regression fit to data from all clips; the dashed line corresponds to an analysis excluding clips where both the human and LENA® said zero child-adult or adult-child turns

seconds to be counted. The association between clip-level LENA® and human CTC was weaker than that found for CVC (Figure 5): the Pearson correlation between the two was r = 0.59 (p = 0.00) when all clips were taken into account, and r = 0.45 (p = 0.00) when only clips with some child speech (i.e., excluding 381 clips with 0 counts in both LENA® and human annotations) were considered.

The *absolute error rate* ranged from -75% to 20%, with a mean of -9.29% and a median of -3%. Excluding clips where both human and LENA® counts were zero, the mean was -16.46% and the median was -13%. Again, *relative error rates* require non-null denominators, and we thus removed the 332 clips in which the human annotator said there were no child-adult or adult-child turns at all. The mean relative error rate ranged from -100% to 400%, with a mean of -67.06% and a median of -83.33%. Finally, the *absolute relative error* rate ranged from 0% to 400%, with a mean of 80.31% and a median of 85.71%. For CTCs, as for CVCs, the LENA® seems to be underestimating counts rather systematically, leading to counts that are on average -67.06% smaller than what they actually are.

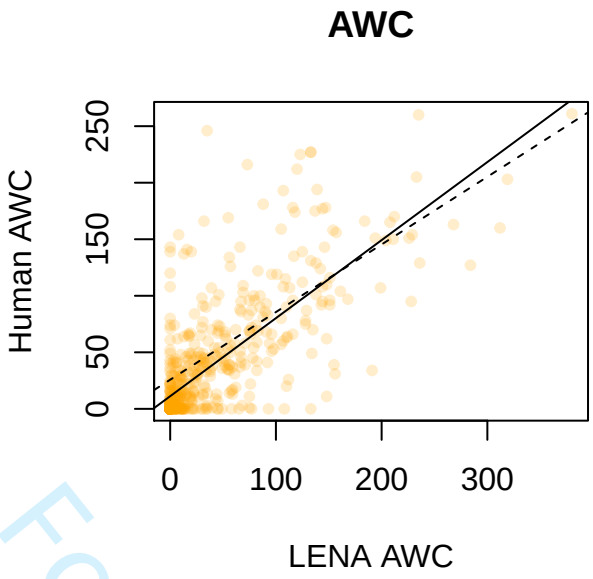

Figure 6. Adult Word Counts according to LENA (x axis) and humans (y axis). Each point represents the AWC totaled within a clip. The solid line corresponds to a linear regression fit to data from all clips; the dashed line corresponds to an analysis excluding clips where both the human and LENA® said there were no adult words.

Adult Word Counts accuracy. The association between clip-level LENA® and human AWC in the four English-spoken corpora was strong (Figure 6): the Pearson correlation between the two was r=0.75 (p=0.00) when all clips were taken into account, and r=0.69 (p=0.00) when only clips with some adult speech (i.e., excluding 309 clips with 0 counts in both LENA® and human annotations) were considered. This suggests that the LENA® system captures differences in terms of number of adult word counts across clips well.

The *absolute error rate* ranged from -211% to 157%, with a mean of -0.20% and a median of 0%. Excluding clips where both human and LENA® counts were zero, the mean was 0.46% and the median was -2%. As for CVC and CTC, calculating *relative error rates* necessitated removing the 361 clips in which the human annotator said there were no adult words whatsoever. The mean relative error rate ranged from -100% to 7400%, with a mean of 55.04% and a median of -17.78%. Finally, the *absolute relative error* rate ranged from 0% to 7400%, with a mean of 123.89% and a median of 58.33%.

Together, these results suggest that the direction of errors for AWC is less stable than in the other two counts (CVC and CTC.) We conclude this from our finding that in the other two counts, the relative error rate and the absolute relative error rate were quite similar, suggesting a stable (underestimation) tendency, whereas here the mean and the median for the relative error rate diverge greatly and both of these are, sign set aside, much smaller than the mean and median absolute relative error rate.

Effects of age and differences across corpora. The preceding sections include results that are wholesale, with all but AWC conducted over all corpora. However, it's possible that performance would be higher for the corpora collected in North America (BER, WAR, SOD) than those collected in other English-speaking countries (L05) or non-English speaking populations (TSI). Additionally, our age ranges are wide, and in the case of TSI children, some of the children are older than the oldest children in the LENA® training set. To assess whether accuracy varies as a function of corpora and child age, we fit mixed models.

We predicted false alarm, miss, and confusion rates in clips where there was some speech according to the human annotatorwhen, with all "F" categories, TV, and overlap were mapped onto silence (which yielded the best results in Section "False alarms, misses, confusion" above.) Our predictors were corpus, child age, and their interaction as fixed effects, with child ID as a random effect. We followed up with a type III ANOVA to assess significance. Across all analyses, corpus, child age, and their interaction were never significant. Supporting this result visually, Figure 7 represents diarization error rate as a function of age and corpus for individual children.

For CVC, we fit a mixed model where CVC according to the human was predicted from CVC according to LENA®, in interaction with corpus and age, as fixed factors, declaring child ID as random effect. This time our Type III ANOVA found a 3-way interaction, suggesting that the predicted value of LENA® with respect to human CVC depended on both the corpus and the child age, as well as all two-way interactions. To investigate this further, we fit this model (removing corpus as a predictor) within each

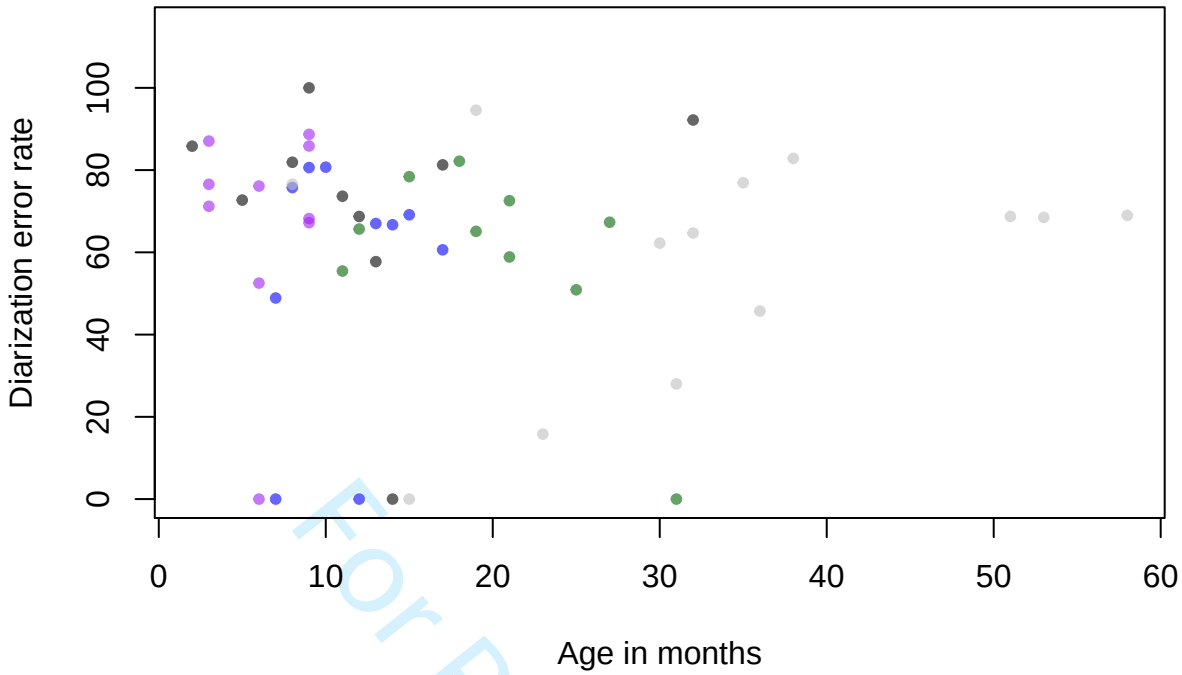

Figure 7. Diarization error rate as a function of corpus and child age. Each point represents the results averaged over all clips extracted from the data of one child. Colors indicate to which corpus that child belongs: BER in blue, L05 is green, SOD in black, TSI in gray, and WAR in purple.

corpus separately. In the SOD corpus, we found a significant interaction between LENA® CVC and age (indicating that the predictive value of LENA® CVC decreased with child age); and a main effect of age (consistent with CVC going up with age in this corpus). The main effect of LENA® CVC, which is consistent with a predictive value of this metric with respect to human CVC, emerged as a significant predictor in all corpora.

For CTC, we fit a mixed model where CTC according to the human was predicted from CTC according to LENA®, in interaction with corpus and age, as fixed factors, declaring child ID as a random effect. This time our Type 3 ANOVA found a main effect of the LENA® CTC estimates and a main effect of corpus, but no significant interactions.

Finally, for AWC (which was only analyzable for the four ACLEW corpora), we fit a mixed model where AWC according to the human was predicted from AWC according to

LENA®, in interaction with corpus and age, as fixed factors, declaring child ID as random effect. The Type 3 ANOVA revealed a three-way interaction, which was investigated by fitting additional mixed models to each corpus separately. An interaction between LENA® AWC and age was found for BER and SOD, due to a *decreased* predictive value of the LENA AWC with respect to the human AWC for older infants in the former but an *increase* in the latter. Notably, the positive association between LENA and human AWC was, significant for all four corpora.

Discussion

The aim of the present study was to assess LENA® accuracy across key outcome measures: speaker classification accuracy, adult word counts, child vocalization counts, and conversational turn counts. We did this using an approach that sought to avoid inflating accuracy estimates in several ways. Methodologically, we used random or periodic sampling to select portions of the files for manual annotation, and our human annotators did not see the LENA® segmentation. Analytically, we considered both speech and non-speech classes (including electronic sounds and silence). This permitted a systematic, extensive, and independent evaluation of key automated metrics provided by LENA®. We also tested generalizability by analyzing the performance of the LENA® system across five different corpora: three based on the same population, language, dialect, and age group that LENA® was established for, and trained on (North American English); one that allowed us to test how accurately it captured a different dialect of English (UK English); and one that tested its performance in a totally different recording situation (a rural setting with large families and many children present, speaking a linguistically unrelated language, and where the key children were, on average, somewhat older). We begin but recapping our key results.

Our first set of analyses tested overall accuracy, using established speech and talker segmentation metrics (false alarm rate, miss rate, confusion rate, and the composite diarization error rate), and evaluated the pattern of errors in more detail, by assessing how

LENA® and human annotators agreed (precision and recall). The median diarization error rate was relatively high (69%), mainly due to a high miss rate (missing or excluding speech that was there; 37%). The false alarm rate (identifying non-speech/silence as speech; 5%) and confusion rate (identifying voice type; 6%) were low.

However, this low overall confusion rate hides considerable errors affecting the interpretation of the output. In terms of precision (to what extent do LENA® tags contain what they say they contain), the system performed relatively well at identifying female voices (69% of frames tagged by LENA® as FAN were coded as female adult by the human coders), and reasonably well at identifying the target child (49% of frames tagged by LENA® as CHN were correct). However, the system performed substantially worse with other talker types (e.g. 9% and 8% for MAN and CXN, respectively); that is, less than a tenth of the frames that LENA® tagged as being speech spoken by these speakers actually correspond to them.

In terms of recall (how accurately LENA® captured the human annotations), performance for the key child's vocalizations was relatively robust; close to half of the frames tagged by humans being key child speech were captured by LENA® under the CHN tag. However, recall was poorer for other talkers: only 28% of adult female near speech (FAN) and less than 20% of adult male and other child speech were correctly tagged by LENA®.

Our second set of analyses tested the accuracy of three of the aggregated counts automatically provided by LENA®, namely Child Vocalization Counts (CVC), Conversational Turn Counts (CTC) and Adult Word Counts (AWC). We found high correlations between clip-level counts estimated via the LENA® system and those from the human annotations for AWC and CVC, with weaker performance for CTC.

However, such correlational analyses do not establish whether LENA® systematically over- or under-estimates. For this we examined several types of error rates. For absolute error rate (i.e. how far was the LENA® count from the human annotators'), the median across clips for CVC, CTC, and AWC was an encouraging 0%,-3%, and 0%, respectively. Notably, this was partly due to many clips lacking vocalizations, turns, or adult words

altogether as without these zero valued counts, the medians were -4%,-13%, and -2%. By and large, LENA® systematically underestimates the raw counts of its main quantitative measures - particularly child vocalizations and conversational turns, and to a lesser extent, adult words, which showed more erratic error patterns.

That said, LENA® results were surprisingly robust to dialect, language, and child age, as we found in our final set of analyses, which tested how well all of these metrics generalized across corpora. We predicted that performance would be higher for the corpora collected in North America (BER, WAR, SOD) than for corpora collected in other English-speaking countries (L05) or non-English speaking populations (TSI), and that accuracy would decrease with age, since our sample contains children older than those included in the LENA® training set. Contrary to our predictions, there were no significant differences in rates of false alarms, misses, or confusion by corpus, child age or their interaction, nor were there differences between North American and other corpora in terms of CTC.

For CVC and AWC, the accuracy of LENA® varied across corpora and age in a way that could not be easily captured by age or dialect. For instance, CVC accuracy decreased with age for SOD whereas it seemed stable in the other corpora. We do not find such cross-corpus differences particularly compelling because there are only 9-10 children in each corpus, and each is typically represented by a single recording, and thus some significant effects may reflect sheer chance. For instance, SOD overlaps in age range with L05 and yet only in the former was an age-interaction with age found, suggesting this result may best be treated gingerly.

More generally, a caveat is in order for all of these apparently null or unstable effects: Absence of a significant difference across corpora and child age may be due to quite considerable *within*-corpus variance combined with relatively small human-coded samples from each corpora. Indeed, further work on bigger samples (more data per child, more children per corpus) is critical to ensure sufficient power and precision.

That said, the current data lead us to conclude that there were very few systematic

language, dialect, and age effects. This is a promising finding for future speech technology solutions, since it suggests that diarization success may not require large quantities of highly specific training data. This is particularly important for the possibility of using LENA® for languages unrelated to the (dialect of) English for which it was developed. That is, given the significant challenges in creating new datasets for training and testing speech technology, if it indeed turns out to be the case that automated speech processing (e.g. diarization) for daylong recordings is relatively language- or dialect-agnostic in its accuracy, it may be possible to use existing American English corpora with high hopes of generalizability.[2]

Summarily, LENA® performs relatively well in terms of overall accuracy, but there are pockets in the system (e.g., identifying male adult voices, establishing absolute numbers of child-adult turns) where the results of the algorithms are quite unreliable. Thus, whether LENA® results are "good enough" for a given research, educational, or clinical study depends largely on the goals of each particular study. For example, we can describe precision rates of 69% (i.e.69% of frames tagged by LENA® as FAN were coded as female adult by human coders) and 49% (i.e.49% of frames tagged as target child were also tagged as such by human coders) as being reasonably good. We do this partly because these rates are much higher than the system's precision rates for other speakers (MAN 9%, CXN 8%) but also partly because our frame-based criteria is more stringent than many coding schemes previously applied. That said, whether a *particular* accuracy rate can be considered sufficient will depend on the purpose of the study. As a result, we next provide a set of recommendations to help researchers make this determination for their goals.

What research goals can one pursue given the performance of LENA® segmentation and metrics? In the present corpora, the system's false alarm rate (i.e., identifying speech where there was none) was very low while its miss rate (missing speech that was actually there) was relatively high. This makes LENA® more suitable for studies in

[2]Please note, however, that since we were unable to compute AWC for Tsimane', we are only able to speak to generalizability across samples and dialects of English for that metric.

which it is extremely important not to "invent" speech that is not there but less suitable for studies in which capturing most, if not all, of the speech produced is crucial. Based on these findings, LENA® would be a good tool for finding "high talk volume" parts of the day for a) careful further transcription (e.g. of low-frequency events like a certain grammatical construction of interest), b) annotation of specific speech characteristics (e.g. mean length of utterance), or c) comparing relative talk volume across samples. However, we advise caution in using LENA® when raw quantity of speech is crucial for the research question, or when small differences in talk volume might have very significant theoretical consequences; this is often the case in clinical populations where children's own vocalizations can be an important diagnosis-relevant characteristic (e.g., in children who are deaf or hard of hearing, individuals with ASD, speech apraxia, etc.).

Similarly, although the overall confusion rate (i.e. incorrectly identifying talkers, such as giving a "mother" tag for a "child" utterance) for LENA® was very low, this does not fully convey the level of accuracy for speech, particularly when considering every talker type. In terms of precision, the system's female adult and key child categorization was quite accurate, whereas precision was lower for male adults and other children, such that the vast majority of the frames labeled as male adult or other children did not in fact contain speech by these speaker types. In terms of recall, LENA® was good at capturing speech by the key child as such, but recall was lower for the other voice categories, with again very poor performance for speech by male adults and non-target children.

We, thus, recommend caution before undertaking any analyses that rely on the accuracy (precision and/or recall) of male adult and other children's speech. For example, if the goal is simply to calculate an overall adult word count (AWC), summing over male and female adult speakers, some confusion between MAN and FAN is likely not problematic. However, if the goal of the study is to compare the relative input from fathers and mothers, LENA® tags are relatively unreliable and on our view, merit further manual vetting in most use cases.

As another example (detailed further in the "Recall" results above), if the goal is to capture as many of the key child's vocalisations as possible, it might be worthwhile to pull out segments LENA® labelled as non-target child, CXN, (of which 33.27% was target child speech) as well, with human coders brought in to filter out non-target child speech. Indeed, we find that this kind of binary classification (key child or not) can be readily undertaken with little training by research assistants in our labs, and would substantially boost data quality and quantity for child vocalizations in this use case.

Notably, while we recommend LENA® users be cautious in their use of LENA® diarization and classification, especially for certain talker classes, our results for LENA® count metrics suggest these derived counts may be accurate enough to serve well across a large variety of uses. To begin with, as far as it is possible to generalize from the limited range of samples tested here (children aged 2 to 58 months, learning North American English, UK English, or Tsimane') it seems that the system's performance does not vary a great deal across ages, dialects, language and home settings. Moreover, *correlations* between human and LENA® clip-level counts were high to very high, suggesting that the software accurately captures differences in counts across clips (even when *error* rates were also high). These correlations remained quite high even when clips with counts equal to zero were removed from consideration, suggesting that LENA® captures gradience in vocalization counts.

However, our finding that LENA® generally underestimates the quantity of child vocalizations, child-adult turns, and adult words deserves further consideration. Indeed, further work is needed to fully understand the nature and extent of this limitation. Our clips were 1–2 minutes in length, and therefore they either tended to have very little speech or a lot of it. Error rates over hours could be smaller, because local errors average out; or greater, if the LENA® system systematically underestimates counts. In a LENA® technical report, AWC accuracy was variable across two 12-hour recordings: 1% lower than human transcription for one child, but 27% lower for a second child. This same report notes that AWC accuracy quickly plateaus as recording time increases beyond one hour, leveling to

5-10% in recordings greater than 2 hours in length (D. Xu et al., 2009).

Thus, it is important for further work to help establish the systematicity in the estimates provided by LENA®: if underestimates are robust and systematic (as suggested by present results for CVC and CTC, but not AWC), it may be possible to develop a correction factor to compensate for this bias. However, this bias may be challenging to nail down precisely and for AWC it may depend on the language in question. For instance, a recent study in Finland documented that LENA® largely *over*estimated AWC and only slightly underestimated child vocalization counts (Elo, 2016).

How to test the reliability of the automated output provided by LENA®. We are overall hopeful that the reliability metrics we provide here will be relevant for researchers working with different populations. We hope the current paper inspires others to evaluate and report all aspects of the system, rather than a subset of metrics. Similarly extensive evaluations of LENA® in other corpora would bolster the validation literature, and be useful for the whole research community. In fact, it would be useful if researchers systematically test the reliability of LENA® counts in their own samples, especially if they are collecting data from families living in different environments from those assessed here. Next, we provide some guidelines for how to go about this. Note that this requires downloading the audio (.wav) file generated by LENA® as well as the concomitant LENA® output file.

First, we recommend a literature search, to determine whether a similar sample has been studied in the past for which there exists reliability data (see for example, A. Cristia et al., under review for a systematic review). If no studies exist, draw 10 x 2 minutes randomly from 10 children. This is about 3h20min of data, which takes roughly 60h to annotate, in our experience. We recommend training annotators using the ACLEW Annotation Scheme https://osf.io/b2jep/, which has an online test annotators can go through to ensure reliability. Once the manual annotations are complete, the LENA® annotations can be extracted and compared against the human annotation using the free DiViMe software

(divime.readthedocs.io, Le Franc et al., 2018). This will allow researchers to extract the classification accuracy measures used here (false alarm rate, miss rate, confusion rate and the derived diarization error rate), as well as CVC, CTC, and AWC comparing LENA® and human annotations. We note this DiViMe pipeline is only possible "off the shelf" for manual annotations made using the ACLEW Annotation Scheme, though in principle, it is adaptable to other schemata by adept programmers.

Separately, researchers should reflect about the accuracy needed for their question of interest. For instance, suppose we have an evaluation of an intervention where we expect treatment children to hear 20% more speech than controls, or an individual differences study where we expect that the lower fifth of the children hear 20% less speech than the top fifth. If the intended measure used to compare groups has an error rate larger than the effect predicted (such as the the CTC error rate we find here), a different algorithm or outcome metric would be wise.

Conclusions

In conclusion, in this study, we have provided a broad evaluation of accuracy across the key outcome measures provided by LENA® (classification, child vocalization counts, conversational turn counts, and adult word counts), and its generalizability across different dialects, languages, ages, and settings. We have provided some recommendations for how to use LENA® in future studies most effectively, and how to test the accuracy of the LENA® algorithms on particular samples of data.

There are, however, a number of areas of research that we have not addressed. For example, we have not investigated how accurately LENA® detects individual variation across children or families. It would be particularly useful to know whether LENA® can classify children with the sensitivity and specificity needed for accurate identification of language disorders. Oller et al. (2010) used LENA® to differentiate vocalizations from 232 typically developing children and children with autism or language delay with a high degree of

accuracy. However, key to this was the use of additional algorithms, not yet available from LENA®, to identify and classify the acoustic features of "speech-related vocal islands". Further work (including shared code) would greatly bolster progress on this topic.

Even if it turns out that LENA® is not accurate enough to classify children precisely for a given ability or diagnosis, it may be accurate enough to capture the rank order of individual children's language growth, which can provide useful information about the relative language level of children in a sample or population (see, e.g., Gilkerson et al., 2017). Similarly, LENA® may not accurately capture the precise number of child vocalisations produced over time, but it may track developmental trajectory (e.g., the slope of growth) relatively well. Finally, although our results suggest that aspects of the system's output may be relatively robust to differences across languages and dialects, we need more evidence of how it fares across mono- and multi-lingual language environments (cf. Orena, 2019).

It is undeniable that children learn language from the world around them. Naturalistic daylong recordings offer an important avenue to examine this uniquely human development, alongside other fundamental questions about human interaction, linguistic typology, psychology, and sociology. Tools and approaches that allow us to tap such recordings' contents stand to contribute deeply to our understanding of these processes. We look forward to further work that addresses the many remaining questions within this area.

Acknowledgments

This research benefits from the Analyzing Child Language Experiences around the World (ACLEW) collaborative project funded by the Trans-Atlantic Platform for Social Sciences and Humanities "Digging into Data" challenge, including a local Academy of Finland grant (312105) to OR, ANR-16-DATA-0004 ACLEW to AC, NEH HJ-253479-17 to EB, and funding from the Social Sciences and Humanities Research Council of Canada (869-2016-0003) and the Natural Sciences and Engineering Research Council of Canada (501769-2016-RGPDD) to MS. AC acknowledges further support from (ANR-17-CE28-0007

LangAge, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the James S. McDonnell Foundation Understanding Human Cognition Scholar Award. MS was also funded by a Social Sciences and Humanities Research Council of Canada Insight Grant (435-2015-0628). EB acknowledges NIH (DP5 OD019812-01). CR was also funded by the Economic and Social Sciences Research Council (ES/L008955/1). OR was also funded by an Academy of Finland grant no. 314602.

Open Practices Statement

The study relies indirectly on daylong audiorecordings, which cannot be made public to protect participants; and human and LENA® annotations for extracted clips, which can be completely deidentified, and which are stored in private repositories which do not have a persistent identifier. These basic data were used to generate statistics at the clip level, which are the input to analyses presented here. Both the latter data and all analyses are publicly available from https://github.com/jsalt-coml/lena_eval. None of these analyses were pre-registered.

References

Bergelson, E. (2016). Bergelson Seedlings Homebank corpus. doi:10/T5PK6D

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724.

Bergelson, E., Cristia, A., Soderstrom, M., Warlaumont, A., Rosemberg, C., Casillas, M., ... Bunce, J. (2017). ACLEW project. Databrary.

Bredin, H. (2017). Pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *INTERSPEECH* (pp. 3587–3591).

Bulgarelli, F., & Bergelson, E. (2019). Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic day-long recordings. *Behavior Research Methods*, 1–13.

Busch, T., Sangen, A., Vanpoucke, F., & Wieringen, A. van. (2018). Correlation and agreement between Language ENvironment Analysis (LENATM) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*, *50*(5), 1921–1932. doi:10.3758/s13428-017-0960-0

Canault, M., Le Normand, M. T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENATM) in European French. *Behavior Research Methods*, *48*(3), 1109–1124. doi:10.3758/s13428-015-0634-8

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., & Sloetjes, H. (2017). A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of children's language environments. In *Interspeech 2017* (pp. 2098–2102).

Cristia, A., Bulgarelli, F., & Bergelson, E. (under review). Accuracy of the Language Environment Analysis System: A systematic review. Retrieved from

https://osf.io/4nhms/

d'Apice, K., Latham, R. M., & Stumm, S. von. (2019). A naturalistic home observational approach to children's language, cognition, and behavior. *Developmental Psychology*.

Elo, H. (2016). *Acquiring language as a twin*. Tampere, Finland: Tampere University Press.

Ganek, H. V., & Eriks-Brophy, A. (2018). A concise protocol for the validation of Language ENvironment Analysis (LENA) conversational turn counts in vietnamese. *Communication Disorders Quarterly*, *39*(2), 371–380.

Gilkerson, J., & Richards, J. A. (2008). The LENA Natural Language Study. LENA Foundation.

Gilkerson, J., Coulter, K. K., & Richards, J. A. (2008). Transcriptional analyses of the LENA natural language corpus. LENA Foundation.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., ... 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. doi:10.1044/2016_AJSLP-15-0169

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., ... Toppings, K. (2016). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech Language and Hearing Research*, *85*(2), 445–452. doi:10.1044/2015

Goh, K.-I., & Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, *81*(4), 48002.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly*, *32*(2), 83–92. doi:10.1177/1525740110367826

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., ... Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics,*

Speech and Signal Processing. Hong Kong.

Le Franc, A., Riebling, E., Karadayi, J., Wang, Y., Scaff, C., Metze, F., ... others. (2018). The aclew divime: An easy-to-use diarization tool. In *Interspeech* (pp. 1383–1387).

Lehet, M., Arjmandi, M. K., Dilley, L. C., Roy, S., & Houston, D. (2018). Fidelity of automatic speech processing for adult speech classifications using the Language ENvironment Analysis (LENA) system. *Proceedings of Interspeech*, 3–7.

MacWhinney, B. (2017). Tools for Analyzing Talk Part 1: The CHAT Transcription Format. Carnegie.

McDivitt, K., & Soderstrom, M. (2016). McDivitt homebank corpus.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... Cutler, E. A. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354–13359. doi:10.1073/pnas.1003882107

Orena, A. J. (2019). Growing up bilingual: Examining the language input and word segmentation abilities of bilingual infants. PsyArXiv. doi:10.31234/osf.io/x9wr8

Rowland, C. F., Bidgood, A., Durrant, S., Peter, M., & Pine, J. M. (2018). The Language 0-5 Project. University of Liverpool. doi:10.17605/OSF.IO/KAU5F

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2019). Second DIHARD Challenge evaluation plan. *Linguistic Data Consortium, Tech. Rep.*

Scaff, C., Stieglitz, J., Casillas, M., & Cristia, A. (in progress). Daylong audio recordings of young children in a forager-farmer society show low levels of verbal input with minimal age-related change.

Seidl, A., Cristia, A., Soderstrom, M., Ko, E.-S., Abel, E. A., Kellerman, A., & Schwichtenberg, A. (2018). Infant-mother acoustic-prosodic alignment and developmental risk. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1369–1380.

Soderstrom, M., Bergelson, E., Warlaumont, A., Rosemberg, C., Casillas, M., Rowland, C.,

. . . Bunce, J. (in progress). The ACLEW Random Sampling corpus.

VanDam, M., & De Palma, P. (2018). A modular, extensible approach to massive ecologically valid behavioral data. *Behavior Research Methods*. doi:10.3758/s13428-018-1167-8

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. In *Seminars in speech and language* (Vol. 37, pp. 128–142). Thieme Medical Publishers.

Warlaumont, A., Pretzer, G., Walle, E., Mendoza, S., & Lopez, L. (2016). Warlaumont HomeBank corpus.

Weisleder, A., & Fernald, A. (2013). Talking to children matters. *Psychological Science*, *24*(11), 2143–2152. doi:10.1177/0956797613488145

Xu, D., Yapanel, U., & Gray, S. (2009). Reliability of the LENATM Language Environment Analysis System in young children's natural home environment. LENA Foundation.

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, *124*(1), 342–349.