BEHAVIOR RESEARCH METHODS

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

Journal:	*Behavior Research Methods*
Manuscript ID	BR-Org-19-300.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	16-Jan-2020
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

Dear Editor,

Thank you for the opportunity to submit a revised manuscript. We have incorporated all suggested changes, which we briefly summarize as follows:

- Added comparisons of the level of agreement for LENA-human against human-human inter-rater reliability, and against state of the art systems
- Added text to recommend caution regarding cross-cultural validity and more generally the follow-up statistical analyses, and removed a paragraph concluding that results were stable across corpora
- Added information regarding interactions
- Created an online supplementary materials website, containing full reports on all analyses (in a pdf format) and all data and scripts to reproduce the present manuscript and that report

We have also made considerable changes to wording, terminology, flow, and focus, to increase clarity and reading ease.

In the process of creating a reproducible pipeline, we found and corrected some errors. These led to minimal changes in the results, none which impacted our interpretation. We thank you for the extension that allowed us to verify this new pipeline.

We hope you will agree with us that the manuscript is an important contribution to this topic, and has an unprecedented wealth of information on the accuracy of LENA algorithms, which our colleagues are likely to find extremely useful.

We look forward to hearing your response.

Alejandrina Cristia, on behalf of all authors

18-Sep-2019

Dear Dr. Cristia:

Manuscript ID BR-Org-19-300 entitled "A thorough evaluation of the Language Environment Analysis (LENA) system", that you submitted to Behavior Research Methods, has been reviewed. The comments of the reviewer(s) are included at the bottom of this letter.

I was able to recruit three excellent reviewers, one of whom was recommended by the authors. I also read the paper with interest myself. I fully agree with the three reviewers that this is an impressive and timely evaluation of a system that is widely used across multiple different fields. And I congratulate the authors for taking on such a difficult and involved project. I also agree with the three reviewers that BRM seems like the right venue for this paper, and that if the authors decide to make a major revision, the paper should eventually be published here. I also share several key concerns about the paper, especially about the evaluation, and the conclusions reached. I simply list my main concerns here, and try not to repeat all the concerns of the reviewers.

1. While reading the paper, my main concern was that comparing the results to manually annotated corpora is great, but **why not also compare to the state of the art audio classification system available? Technology, and specifically, machine learning, has changed dramatically since LENA was developed. If we want to answer whether LENA is a reliable research tool, we need to compare it to the state of the art systems. Both the second and third reviewer agree with this assessment. What does it mean that LENA can only identify a target child only at 50%? We need to know how 2-3 other state of the art systems perform and compare LENA's performance to those. Reviewer 2 and 3 provide citations to some papers for this**.
RESPONSE: We have added references to the DIHARD challenge, which was partially based on comparable data. Our reanalyses of the submitted systems shows Diarization Error Ratess varying between 48% and 121%, with a median around 70%.

2. **The authors cannot make any cross-cultural generalization based on a corpus from only a single other culture which has N=10. Either expand your corpora, or completely tune down the cross cultural claims.**
RESPONSE: We have toned down cross-cultural claims.

3. **What LENA tries to accomplish might even be hard for humans. How well are your annotators doing on the task? We need to know a measure of inter-coder reliability. You are taking human annotations as the gold-standard, without telling us how well the humans do on the task. A nice recommendation by Reviewer 3 is to also report the inter-coder reliability between LENA and the human annotators.**
RESPONSE: We now report reliability on a partial sample.

4. You need to have a more thorough discussion of your analyses. **For example, you report a three-way interaction. What does this interaction capture? Same with other analyses.**
RESPONSE: We have explained the results more clearly, adding discussion throughout..

5. **The samples need to be explained in more detail.**
RESPONSE: We have added detail on the samples.

Reviewer 1 also questions some field-specific questions, all of which should be addressed before this revision.

Please note that if you decide to go for a revision, and I encourage you to do so, **provide a point by point response to all the concerns brought up by the reviewers and myself**. There are also a number of typos through out the manuscript, some of which have been pointed out by the reviewers, and some that were not (e.g. page 6 "in a pee-reviewed journal"). Please make sure you address these smaller points as well.
RESPONSE: We are sorry for those typos. We have made our best to make sure none remained. (And we've corrected the one you point out in this paragraph, though it did make us giggle.)

If you are unable to enter your revision within the **next 90 days**, please email journals@psychonomic.org for an extension.

To revise your manuscript, log into https://mc.manuscriptcentral.com/brmic and enter your Author Center. You will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions" click on "Create a Revision." Your manuscript number has been appended to denote a revision.

You will not be able to make your revisions on the documents previously submitted. Instead, revise your manuscript using a word processing program, highlight the changes, and save it on your computer. (You can highlight the changes within the document by using the track changes mode in MS Word or by using bold or colored text.) Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.

***IMPORTANT: Please be sure to enter your response to the editor's and reviewer's concerns in the author's response field, not as a separate cover letter or attached file. Your original files are carried into the revision. Please delete any redundant files (such as the earlier cover letter, manuscript, or figures no longer the same) before completing the submission of your revision.

Because we are trying to facilitate timely publication of manuscripts submitted to Behavior Research Methods, your revised manuscript should be uploaded as soon as possible. If it is not possible for you to submit your revision in a reasonable amount of time, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to Behavior Research Methods and I look forward to receiving your revision.

Sincerely,
Morteza Dehghani
Action Editor, Behavior Research Methods

Reviewer(s)' Comments to Author:
Reviewer: 1

Comments to the Author
Thank you for doing this work. Proper validation of these measures is vital to the future of automated vocal analysis tools. This paper validates the LENA System using a variety of

comprehensives measures and diverse populations using a common protocol for human coding. While the authors were able to demonstrate that the device is relatively accurate there were segments that were less easily identified. They go on to provide recommendations about how and when clinicians and researchers can use the LENA System. Their synthesis makes implementation of their results practical and manageable. Stakeholders should easily be able to interpret the findings from this paper to improve their LENA use and, in the future, use of daylong recordings for research and clinical purposes. To that end, I have a few small points of clarification I believe will help readers best understand the work presented here.

Page 8: Why did you choose not to focus on TVN?
RESPONSE: Our primary interest is in children's language development, for which TV is not supposed to provide the most critical input. Thus, TVN is not a crucial category, from our point of view. We have not modified the thrust of manuscript to address this question.

The human coding protocol evaluated 10ms frame. How does that equate to the amount of time the software needs to label a segment? What is the shortest frame the LENA System can label?
RESPONSE: We have rewritten the section that probably triggered these questions as follows: *"the fact that annotators did not have access to the LENA^®^ segmentation allowed an assessment of the accuracy of the segmentation itself as well as categorical labeling. Specifically, LENA^®^ and human levels were evaluated every 10 ms. 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)."*

Before, we were giving the impression that we chose this frame size in order to have a finer-grained view, whereas the reality is we need some frame size because our annotators do not have access to LENA's segmentation. Some information regarding minimal segment duration was already present in the section "Brief introduction to LENA products": *"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; @Xu2009a]. Silence may be included to "pad" segments to this minimum duration."*

We also mention this in the results: *"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)."*

The TSI cohort has older children in it. Is there research showing that the tool can be reliably used with that age range? What is your hypothesis?
RESPONSE: We added our expectation that accuracy may decline due to a mismatch between the LENA training and this sample, as well as references to previous work looking

at accuracy in older children. Specifically, we now say: *"By and large, one could expect accuracy to decline in the sample of children who spoke a different English dialect compared to the three samples that matched better the data the LENA^®^ software was developed with; and one could predict an even greater reduction in accuracy for the group that is learning a completely different language and which further mismatches in age [see other work on age- and language-mismatching samples, @Busch2018; @Canault2016]."*

Page 10: How did you suspect the differences in sampling procedure between the TSI cohort and ACLEW cohorts would influence your results?
RESPONSE: We did not expect this to influence our results in any way. We have not added any text addressing this in the paper, but we can if the Reviewer feels strongly about it.

Page 12: As of yet there is no standard language, to my knowledge, for the labeled time periods that are the subject of these automated vocal analysis papers. Therefore, it is important to be clear about the terms that you will be using. I found the definitions on page 12 and the visual in Figure 1 helpful. However, **in Figure 1, labeling 'segment' and 'frame' within the image would make the concept even more understandable.**
RESPONSE: We have added segments to the Figure, but not frame because it would be invisible in this scale. We have also added this text to the Figure caption: *"Notice that there are multiple clips extracted from each recording; each clip can have zero or more segments; frames(10 ms) are not shown because they would be too small in this scale."*

Page 15: The ACLEW category 'silence' includes sounds from electronic devices and frames the LENA System should classify as OLN. It is clear that this was done to avoid confusions but **the label 'silence' makes it more difficult to understand. Perhaps consider changing it to another term**.
RESPONSE: We have changed this category's label to "Other".

Page 16: How commonly used are the measures you chose? Do other LENA validation studies use them? Should more studies be using them?
RESPONSE: Given the page number, these questions refer to the error rates. Error rates are commonly used in speech technology, and they are not uncommon even in LENA evaluation. In a recent systematic review of AWC evaluations, 8 studies reported both correlations and error rates, 5 reported only correlations, and 6 reported only error rates. We believe correlations and error rates provide complementary information, and thus it would be useful to report both. That said, error rates can be calculated in many ways (as we illustrate), so it would be ideal if authors shared their data so that one can check error rates in multiple ways. We have not added this information to the manuscript

Page 30: line 32, add a space between 'annotator' and 'when'
RESPONSE: Fixed.

Page 31: Please use a **consistent formant for 'Type 3 ANOVA,' either 3 or III**.
RESPONSE: Fixed, sorry about that.

Page 33: **Some of the precision and recall percentages strike me as low (eg. 49%) but the authors state that this is a good result. What percentage should LENA users be hoping for/accepting of? Are there comparable system that could be used as a fair comparison?**
RESPONSE: This relates to other reviewers' comments about comparing LENA against state-of-the-art systems. We ask R1 to look at our replies to R2 and R3 for further detail, but in a nutshell, in the manuscript, we have added references to the DIHARD challenge, which was partially based on comparable data. Our reanalyses of the submitted systems shows Diarization Error Rates varying between 48% and 121%, with a median around 70%, and thus comparable to the LENA performance.

Reviewer: 2

Comments to the Author
This manuscript assesses the accuracy of the LENA system for recording child language. It does so by directly comparing hand coded estimates of key categories (e.g. speaker, turn count) to those produced by the LENA system. The authors convincingly argue that this work is needed, because the LENA system is commonly used but prior evaluations of its effectiveness have key limitations: validation of the LENA method was not the main goal of the review (e.g. it was part of a supplement, and so may not have gotten sufficient attention), incomplete validation information was often reported (e.g. only codes relevant to the empirical claim), and the LENA system itself was used to select a subsample of clips which were to be validated. The authors also collected data on children from different regions (e.g. US, UK, Tsimane), which improves the generalizability of the assessment of the research tool. This is exactly the kind of research question that I would be happy to see addressed in this outlet.
However, there were some limitations to the manuscript that made me believe this question had not yet been adequately addressed.
First, the conclusion in the manuscript was that "LENA performs relatively well in terms of overall accuracy" (p. 35, line 20). This does not seem to reflect the evidence conveyed. The medina Diarization Error Rate was 69%. Other papers assessing DER report much lower error rates (e.g., Barras, Zhu, Meignier, & Gauvain, 2017, report 8.5%; Wooters & Huijbregts, 2008, report 8.5%), and the Google AI blog reports 7.6%. While the authors make the point that identification of female and child voices were relatively better than other speaker types, this error rate is much higher than the current state of the art. I understand that the authors are not presenting a new algorithm, but I think the claim that LENA performs well in overall accuracy needs to address the existing benchmarks in the literature.
RESPONSE: We thank the reviewer for bringing up this important topic. As noted by Mighafori & Wooters (2006), in relation to files with unusually high DER: "*[such files] tend to have many speakers, a large number of speaker turn changes (and therefore, short turn durations and a high turns-per-minute rate) and a high do-nothing DER (i.e., the dominant speaker is not voluble). The correlation of these factors with high DER is compelling, as long uninterrupted speech segments spoken by only a few speakers seem intuitively easier to diarize than frequently interrupted short segments from many speakers where no one*

speaker is dominant." This is exactly the situation we find ourselves in with daylong recordings: many speakers, short turns, and lots of non-speech. Thus, comparing our DER to the references from R2 and R3 is a bit problematic, as detailed below.

- **Barras, Zhu, Meignier, & Gauvain, 2017, report 8.5%: (**we believe R2 meant 2004 as the year, i.e. this paper). This impressive result is based on broadcast news, which reflects a very formal speaking style in a carefully-recorded low-noise setting with few speakers. That is, the recordings that this system was tested on are far less likely to feature multiple overlapping talkers, background media, toy sounds, rustling, car/street noise, etc. than our daylong naturalistic recordings. Moreover, unlike day-to-day conversations, newscasters are far more likely to monologue, or converse in organized turns. and speak in complete sentences relative to spontaneous natural conversation.
- **Wooters & Huijbregts, 2008, report 8.5%.** Here too there are some large differences between the data used for dev/eval and our daylong child recordings. Firstly, the 8.5% DER R2 notes is for the multiple distant microphone condition; in the single distant microphone condition the authors report ~2.5x greater DER of 21.7%. The authors call this gap 'striking'. Our recorders have a single microphone. Similarly to our description of the news data set in Barras et al, here too the data are meetings, which have a more formal style and lower speaker-count than naturalistic daylong recordings.
- **Google AI blog reports 7.6%.** Here R2 refers to a blog referencing Zhang et al 2019 (the same references noted by R3). Alongside the data-type differences highlighted above that we highlight here as well, we note that this work omitted all overlapping speech from evaluation, permitted 250ms of error in segmentation, again making it quite incomparable to our approach.

In sum, previous work has looked at easier data sets, and furthermore inflates performance reports by adopting lenient evaluation schemes. The only comparable data point in the speech technology literature comes from the DIHARD Challenge. DIHARD employed data from a range of domains, including daylong recordings; in fact, they used a different selection of data from the BER corpus used here. The subset of BER used for DIHARD is likely to lead to lower error rates because they selected only files that contained some speech; by excluding files with little to no speech, they prevent the appearance of very high DERs (which emerge when the numerator, i.e. the amount of speech, is very small). Thus, we should expect the DIHARD reanalyses to lead to an *over*estimate of systems' performance. When we reanalyze the DIHARD data for BER, we find DERs varying between 48% and 121%, with a median around 70%.

The sample should also be addressed in greater detail. **Given the relative difference in performance in identifying male versus female speakers, I wanted to see the number of male adult and female adult speakers in each sample.** My impression was that there were few men in the sample, and this makes me skeptical that we should draw any conclusion about the ability of this study to address accuracy of male speech identification.
RESPONSE: We have

- highlighted information on the proportion of frames human annotators attributed to each speaker category, which was already present, by showing it in a table format, and adding the number of frames and minutes each corresponds to. The Reviewer's

comment seems to talk about the number of different people within each category (number of different males) but LENA does not distinguish individual speakers. Thus, the proportion of data of each kind is more relevant.

- compared the prevalence of males and other children against previous evaluations in this response (we did not include this information in the paper, but we can, if the reviewer believes it would be desirable. Out of 32 evaluations, 4 provide data on male prevalence and 2 on other children's prevalence in an interpretable way (that is, some papers report the number of samples containing male speech and those containing female speech, but none of the other categories, so we do not know what is the total number of samples inspected). For Male Adult prevalence, the percentage of males out of all coded samples is 9.4% for Bergelson, Casillas, et al. 2019, who used a method that prioritized getting speech samples, whereas for most others it is much lower: Lehet et al. 1.4%; Soderstrom 2016 1.4%. For other children, the prevalence is 15% for Soderstrom 2016 (likely because two thirds of the children were recorded at daycare, rather than in their homes). Bulgarelli & Bergelson 2019 do not report proportion of vocalizations but rather proportion of coded nouns; they find 17% of the nouns were spoken by males, and 5% by other children.
- added some text in the discussion on how low prevalence of a class affects training (it does - which is why many approaches include data augmentation and resampling) and how it effects validation (it typically does not; provided this represents prevalence in target uses). Specifically, we say:

"One issue that may arise is whether data should be sampled differently to, for example, make sure every class is represented the same amount of time and/or a minimum of time. Our understanding is that class imbalance and data scarceness is an important issue for training, and directly affects algorithm accuracy (this is a general problem, but to cite just one example on GMMs, Garcia-Moral, Solera-Urena, Pelaez-Moreno, & Diaz-de-Maria, 2011). However, it does not pose the same kind of problem for evaluation. That is, if there are no samples of a given category, then accuracy cannot be evaluated; if there are only a few, then it is possible that these are special in some way and accuracy estimates may not generalize well to others. Thus, it would indeed be desirable to have enough samples of a given label to reduce the impact of each individual instance, in case they are outliers. That said, almost any strategy that attempts to boost the frequency of specific categories risks worsening non-generalizability concerns. For instance, if one were to over-sample regions tagged by LENA® as MAN in the hopes of having more male samples, one may only be capturing certain types of male speech or acoustic properties. To take this example further, notice that male speech is our smallest category, representing 3% of the data. Since we sampled randomly or periodically, this represents the prevalence of male speech and the samples that are included are unlikely to be acoustically biased."

Additionally, given my background in psychology, **the sample size of 9 or 10 children from each corpus seemed too small to draw particularly strong conclusions about the generalizability of LENA across cultures.** While I applaud the inclusion of Tsimane sound files, I also am not convinced that these 10 children (even given the 272 clips coded) are enough to address a stable generalization. Ideally more individuals would be sampled, but at least the strength of the language could be weakened.

RESPONSE: We have Added text to recommend caution regarding cross-cultural validity and more generally the follow-up statistical analyses, and removed a paragraph concluding that results were stable across corpora

This connects to the broader issue of the tests used to assess differences between samples. This section seemed under-reported, especially compared to the prior sections. For example, no F-statistics, df, or p-values were reported for any tests. The final paragraph of the section labeled “Effects of age and differences across corpora” mentions that a complex 3-way interaction was found and follow-up tests were conducted but fails to explain anything further. What was the interaction? What were the coefficients? What were the follow-up tests? Plotting this and more of this section would also be appreciated.
Additionally, the description of the results overall was difficult to follow. I would suggest structuring the results around the questions being addressed by each analysis and including more summary tables. Thorough reporting is to be encouraged, but more signposts and interim summaries would be helpful for readers.
RESPONSE: We have:

- restructured the “Effects of age and differences across corpora” section of the results around questions
- added a table reporting Chi-squares, df, p for all tests. We do not report coefficients because it would be necessary to add one for each level, which would mean a table with 8 more lines (5 levels of corpora mean 4 coefficients, one for each corpus that is not the baseline, for the main effects; and as many for the Corpus*Age interaction terms). However, this information can now be found in the Supplementary materials, which have a print out of all test results.
- we have added information on the 3-way interaction, including the follow-up tests and their results, but not added a figure representing this three-way because it is difficult to represent. We also want to advise extreme caution in interpreting these interactions given precisely the issues in power and representation the reviewer has brought up (regarding cross-cultural validity, but it is actually generally relevant)

The language surrounding absolute error rate was also confusing. When the metric is introduced, it appears that this is calculated by the formula NL-NH (number of LENA codes minus number of human codes) **averaged across a set of files**.
RESPONSE: We have corrected this language, it is not averaged across a set of files, but instead done within each file: “*It is calculated as NL-NH, where NL is the number according to LENA^®^ and NH is the number according to humans;* ***this is done separately for each clip****. By* ***then*** *averaging across clips, we then get an idea of the bias towards overestimation”*

This would yield a numeric value, not a proportion value. The example of an absolute error rate on p. 16 (line 34) is “-100 vocalizations off (for the absolute error rate). However, the results section reports it with a % sign (see p. 26, line 55, p. 27, lines 30-32, p. 28, line 40, etc.). This may be a simple typo, but given that there are so many different evaluations—some in a raw metric, some in a proportion metric—it led to some real

confusion when reading the manuscript. It is particularly confusing on p. 29 when the reader encounters 7400% error on one metric.
RESPONSE: Sorry that this was unclear: there were error estimations that are full numbers and others that are proportions/percentages with no bound (hence 7400%). For increased clarity, we now report all error rate analyses in tables rather than text.

Overall, I think this is an important question, but I do not find that this manuscript has yet convincingly answered it. I hope that further work on this project will help address the current limitations.

Reviewer: 3

Comments to the Author
I would like to first thank you all for putting this much time, energy, and care in this work. I truly enjoyed reading it and I'll do my best to contribute in a small way as a reviewer.

Short Summary:
This paper compares the performance of LENA's speech annotation algorithms to human annotators on key measures of speaker classification, adult word counts (AWC), child vocalization counts (CVC), and conversational turn counts (CTC), using a set of audio clips extracted from corpora of American English learning children, British English learning children; and Tsimane'-learning children in rural Bolivia. The paper reports "reasonably high accuracy" in AWC and CTC, but problematic levels of performance in CTC. In addition, error rates did not differ significantly across corpora.

Main Comments:
My main "main comment" is this: **The paper sets out to evaluate LENA but it was not clear against which benchmark. Looking at the results, I'm slightly concerned by the extremely high error rates. I personally found the results not satisfactory but the authors seem to have a much more generous take on it. I may be missing something here, but for example, is identifying the target child 50% of the time really a "reasonably well" performance? Again, compared to what benchmark? Is LENA's performance comparable to industry standards?** Would it be possible to include results on state-of-the-art systems in this paper? For example, Zhang et al (2019) (https://arxiv.org/abs/1810.04719) are reporting extremely lower rates (under 10%). Are there reasons to accept such high error rates in LENA? If not what can LENA do, drawing insights from current state-of-the-art systems to improve? In short: in the current version of the paper it is not clear what is LENA being evaluated against.
RESPONSE: We thank the reviewer for raising this important point. As noted by Mighafori & Wooters (2006), in relation to files with unusually high DER: "*[such files] tend to have many speakers, a large number of speaker turn changes (and therefore, short turn durations and a high turns-per-minute rate) and a high do-nothing DER (i.e., the dominant speaker is not voluble). The correlation of these factors with high DER is compelling, as long uninterrupted speech segments spoken by only a few speakers seem intuitively easier to diarize than*

frequently interrupted short segments from many speakers where no one speaker is dominant." This is exactly the situation we find ourselves in with daylong recordings: many speakers, short turns, and lots of non-speech. Thus, comparing our DER to the references from R2 and R3 is a bit problematic, as detailed below.

- **Barras, Zhu, Meignier, & Gauvain, 2017, report 8.5%: (**we believe R2 meant 2004 as the year, i.e. this paper). This impressive result is based on broadcast news, which reflects a very formal speaking style in a carefully-recorded low-noise setting with few speakers. That is, the recordings that this system was tested on are far less likely to feature multiple overlapping talkers, background media, toy sounds, rustling, car/street noise, etc. than our daylong naturalistic recordings. Moreover, unlike day-to-day conversations, newscasters are far more likely to monologue, or converse in organized turns. and speak in complete sentences relative to spontaneous natural conversation.
- **Wooters & Huijbregts, 2008, report 8.5%.** Here too there are some large differences between the data used for dev/eval and our daylong child recordings. Firstly, the 8.5% DER R2 notes is for the multiple distant microphone condition; in the single distant microphone condition the authors report ~2.5x greater DER of 21.7%. The authors call this gap 'striking'. Our recorders have a single microphone. Similarly to our description of the news data set in Barras et al, here too the data are meetings, which have a more formal style and lower speaker-count than naturalistic daylong recordings.
- **Google AI blog reports 7.6%.** Here R2 refers to a blog referencing Zhang et al 2019 (the same references noted by R3). Alongside the data-type differences highlighted above that we highlight here as well, we note that this work omitted all overlapping speech from evaluation, permitted 250ms of error in segmentation, again making it quite incomparable to our approach.

In sum, previous work has looked at easier data sets, and furthermore inflates performance reports by adopting lenient evaluation schemes. The only comparable data point in the speech technology literature comes from the DIHARD Challenge. DIHARD employed data from a range of domains, including daylong recordings; in fact, they used a different selection of data from the BER corpus used here. The subset of BER used for DIHARD is likely to lead to lower error rates because they selected only files that contained some speech; by excluding files with little to no speech, they prevent the appearance of very high DERs (which emerge when the numerator, i.e. the amount of speech, is very small). Thus, we should take the DIHARD reanalyses to lead to an overestimate of systems' performance. When we reanalyze the DIHARD data for BER, we find DERs varying between 48% and 121%, with a median around 70%.

* Is there any **assessment of inter-rater reliability among human annotators**? The paper does not report on it but it would be quite helpful to have some comments on that.
RESPONSE: We have added Cohen's kappa from reliability coding that is ongoing in the ACLEW project: Cohen's kappa agreement was .64 (weighted kappa .65).

* Would it be reasonable to **"also" report some agreement metrics such as Cohen's kappa (human-LENA agreement)**? My reasoning is that, if we assume that LENA is an annotator (possibly replacing humans), then it may be reasonable to apply the same

standards we apply to human annotators for measuring reliability? And if my understanding is correct, Cohen's kappa is going to be much more strict than some of the measures reported in the paper. Although I understand that the paper is following common practices in computational literature in reporting performance of classification systems.
RESPONSE: We have added Cohen's kappa between LENA and human annotators, which was about .43 to .46, somewhat lower than human-human.

Presentation:
* **The results in the section titled "LENA® classification accuracy: False alarms, misses, confusion." on page 18 can be summarized in one table to allow easy comparison across the several ways the data were analyzed.** Reading percentages within a text is difficult.
RESPONSE: We have replaced the text with a table.

* page 13 line 4: **could you explain why 10ms frames?** My own reaction was "well, why not!" but in case you have any reasoning behind it, it might be good to explain it here
RESPONSE: We have rewritten this text as follows: "*the fact that annotators did not have access to the LENA^®^ segmentation allowed an assessment of the accuracy of the segmentation itself as well as categorical labeling. Specifically, LENA^®^ and human levels were evaluated every 10 ms. 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).*"
In a nutshell, we need **some** frame size because our annotators do not have access to LENA's segmentation so we need to match up sections across the two types of annotations somehow. Another option would have been to use larger frames, say 100ms. However, both LENA annotations and human annotations could start in the middle of such a long frame; for instance, at 155ms from the beginning of the file. In that case, we would have needed to make some decision like "this frame will be named on the basis of the majority class" - which still left the problem of what to do when a frame contains exactly half of two classes. In contrast, no LENA annotation can start in the middle of a 10ms frame because the LENA always reports segmentation at 2 decimals of a second, or 10ms, given that this is the size of their feature extraction window. We have not added this to the manuscript.

* **The categories in the Figure 2+3 confusion matrices do not match and more importantly they are not in the same order.** On the x-axis has female, child, male but the y-axis has child, female, male, … . Is there a particular reason for that? I'm used to confusion matrices with the same number of categories on each axis (I understand why you don't have that) and a diagonal showing where high correlation or percentage agreement is to be expected ideally.
RESPONSE: We have reordered the categories such that homologous categories are in the same order, and highlighted the "diagonal".

* May be better to show p < threshold rather than p=0.00
RESPONSE: We have fixed this.

* Figure 7 could use a legend for the color fills and corpora
RESPONSE: We have added a legend.

Computational Reproducibility:
* The github repo is quite nicely organized and I managed to clone it and figure out what's where but it would be helpful to add some instructions and guides in the README file to show how one should go about reproducing this work.
RESPONSE: We have added a README with detailed instructions. Please note that full reproducibility is only possible for people who have access to our cluster because the earliest stages of processing are based on private data. This is indicated in the README.

Typos and such:
* Page 3: line 55: "previous validation" has an extra "
RESPONSE: We have fixed this

* Page 14, line 52: correctly classified "by" LENA as containing talk?
RESPONSE: We have fixed this

* Page 20, line 44: Capitalize "The" in the beginning of "the LENA system's false alarm"?
RESPONSE: This sentence has been removed in the revised submission due to rewording.

* page 30, line 33: annotatorwhen
RESPONSE: We are sorry for those typos. We have made our best to make sure none remained. (And we've corrected the one you point out here.)

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, Child Vocalization Counts (CVC), Conversational Turn Counts (CTC), and Adult Word Counts (AWC). 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, CVC), with more problematic levels of performance in others (CTC, precision of male adults and other children). Statistical analyses do not support the view that performance is worse for children who are dissimilar from the LENA® original training set. 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, 2019). The work undertaken thus far also has some limitations, which are described further in the "Previous Validation" section below. Bearing these in mind, we

endeavored 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 large sample of diverse infants, including (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 socio-cultural 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, with an additional 60 minutes from 70 additional recordings for testing it (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; i.e. 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., 2019) 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 a mean RER [what we call 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."

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)

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 peer-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 or considered to be 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., 2019, 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 temporal 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, also considering non-speech labels (notably silence, overlap, and TV, with limitations in our approach to be discussed below); as well as the system's key derived metrics: Child Vocalization Counts (CVC), Conversational Turn Counts (CTC), and Adult Word Counts (AWC). 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, the fact that annotators did not have access to the LENA® segmentation allowed an assessment of the accuracy of the segmentation itself as well as categorical labeling. Specifically, LENA® and human annotations were compared every 10 ms. 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. By and large, one could expect accuracy to decline in the sample of children who spoke a different English dialect compared to the three samples that matched better the

data the LENA® software was developed with; and one could predict an even greater reduction in accuracy for the group that is learning a completely different language and which further mismatches in age (see other work on age- and language-mismatching samples, Busch et al., 2018; Canault et al., 2016).

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 us better infer the likelihood with which our results will generalize to other corpora, provided the annotation scheme is conceptually comparable.

Methods

This paper was written using RMarkDown (Baumer, Cetinkaya-Rundel, Bray, Loi, & Horton, 2014) in R (Team & others, 2013) running on Rstudio (RStudio Team, 2019). It can be downloaded and reproduced using the data also available from the Open Science Framework, https://osf.io/zdg6s. These online Supplementary Materials also include a document with the full output of all models discussed here as well as additional analyses.

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., 2019); 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"; Scaff, Stieglitz, Casillas, & Cristia, 2019). Socioeconomic status varies both within and across corpora. Key properties of the five corpora are summarized in Table 2.

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	13	272	60	34 [15-58]	Northern Bolivia

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 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 daylong file, fifteen 2-minute non-overlapping sections of 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 13 children, out of the 25 children recorded from field work that year; the other 12 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/vocalizations (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 5-15 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 several 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 (Bergelson et al., 2019 for the general annotation protocol; cf. Casillas et al., 2017; Soderstrom et al., 2019, for an introduction to the databases). The ACLEW datasets also had other coding levels that will not be discussed here. 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.

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 CVC and CTC can be established. In addition, since most previous work evaluating AWC did so at the clip level, we do so here as well.

The other metrics require a more detailed explanation, conveyed graphically in Figure 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 LENA® system's automated analysis, only one of these categories may be active at a given point in time. In contrast, colloquially, "utterance" or "vocalization" refers to stretches of speech detected by humans and assigned to different talkers. To be clear: in what follows, clips may have zero or more utterances. Unlike in the LENA® system, however, in the human annotation 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

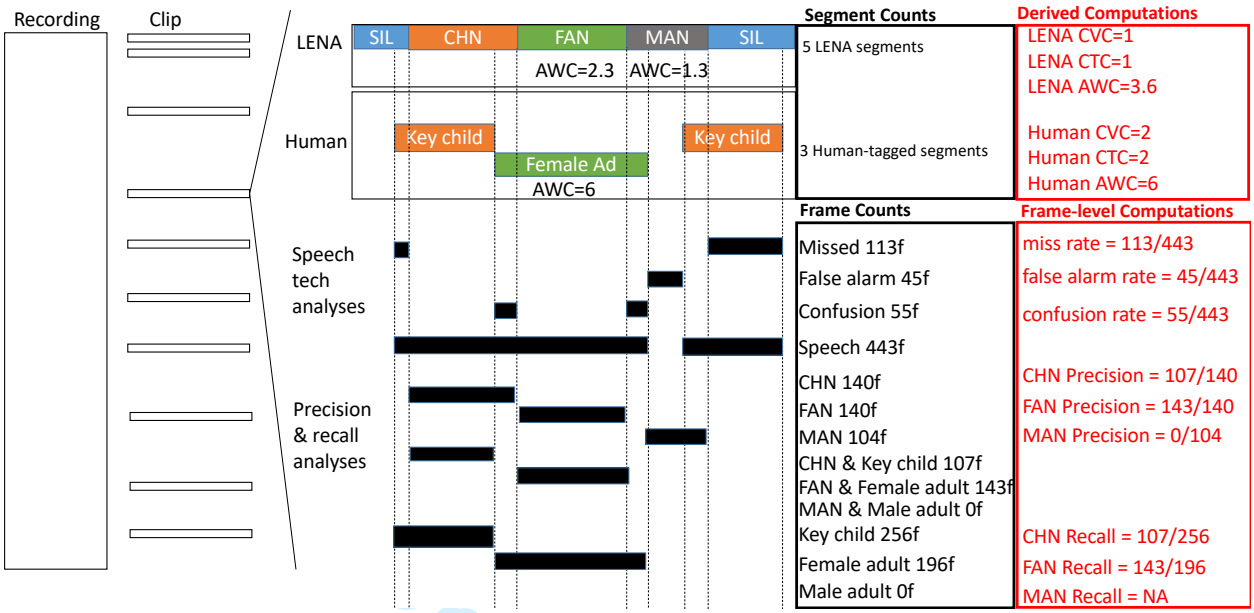

Figure 1. Levels at which performance is evaluated. Notice that there are multiple clips extracted from each recording; each clip can have zero or more segments; frames (10ms) are not shown because they would be too small in this scale. Adult Word Count (AWC), Child Vocalization Count (CVC), and Conversational Turn 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 extracts."f" above indicates 10-ms frames. N.B. for example's sake we assume each child vocalization above has a single linguistic vocalization.

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), while taking into account other categories (with some limitations on their interpretation clarified below). 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 Identification 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.

Since these metrics establish errors relative to speech quantity, a problem emerges when there is no speech whatsoever in a given file. This is never discussed in the speech technology literature, 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 because this was central for our research goals: We need systems that can deal well with long stretches of Other (i.e., non-speech or silence), because we want to measure in an unbiased manner how much speech (and silence!) children hear. Unfortunately, in the 30% 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. 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. Notice that 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, resulting in what may be outliers.

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 identification metrics: rate of false alarm for speech, rate of misses for speech, rate of confusion between talkers, and the derived identification error rate (IDER). 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 by 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:

- **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,
- **ID**entification **E**rror **R**ate (IDER) = (FA+M+C)/T

In the human annotation, there is no class representing overlapping speech as such. For the sake of completeness and greater comparison with the LENA® model, if two or more different 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 LENA® system's OLN (and, for the precision/recall analysis introduced next, OLF).

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 + CHN would be counted towards overlap as would FAN+FAN; whereas for us, only overlap between two sources (e.g., key child and female adult, key child and electronic speech; but not key child + noise since noise was not coded) counts as overlap. Similarly, the TVN LENA® class is not

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	FEM
Male Adult	MAN	MAL
Electronics	TVN*	ELE*
Overlap	OLN	OVL

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, Table 3 mentions some correspondances, but since these are not perfect, additional analyses map overlap and electronic classes onto "Other" post hoc, so as to not penalize LENA® due to a divergence in coding criteria.

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.

Agreement.

When two or more annotators provide data on the same classification, one can calculate agreement. We report on Cohen's κ as a measure of the extent to which LENA® and human annotators coincide in their labeling.

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 if and only if the vocalization had been annotated as being linguistic (canonical or non-canonical in the ACLEW notation).[1] 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 then calculated.

Users may also wish to interpret the actual number of vocalizations or turns found by LENA®. Therefore, it is important to also bear in mind errors, error rates, and absolute error rates. Despite the similarity in their names, these three metrics provide different information. We define *error* as follows: given a LENA® estimate, how close the human-generated value is. This is calculated as NL-NH, where NL is the number according

[1]In a previous version of this analysis, we had calculated CVC as the number of CHN segments in LENA®, and the number of linguistic vocalizations as tagged by human annotators. Further inspection of LENA® documentation revealed this was incorrect, since LENA® counts can include several linguistic vocalizations within one CHN segment, and also includes linguistic vocalizations from CHF segments. Given the inaccuracy of CHF, the latter decision seems potentially problematic. The same issue affected our CTC analyses. We now present analyses here that correctly represent LENA®'s reported CVC and CTC, since these are the field-standard measures. In Supplementary Materials (https://osf.io/zdg6s), we show results of the correlations and error analyses when CVC and CTC are calculated as the number of CHN/CHI segments instead. For CVC the results are identical; for CTC results were slightly worse results than those reported here.

to LENA® and NH is the number according to humans; this is done separately for each clip. 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).

In contrast to *error*, *error rate* computes 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 according to the average error; 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 error rate, we can check whether this seemingly small difference is indeed small relative to the actual number found. That is, an error of 10 vocalizations would be less problematic if there were 100 vocalizations on average (in which case LENA® would be just 10% off) than if there were 10 (LENA® would be doubling the number of vocalizations). As with error, the sign of this difference indicates whether LENA® tends to over- or under-estimate these counts.

Finally, the *absolute error rate* is calculated with the formula abs(NL-NH)/NL, where abs indicates absolute value. 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 an error of 0 together with an error rate of 0 because half of the samples are -100 vocalizations off (for the error) or -100% off (for the error rates), with the other half behaving in the exact opposite fashion. The *absolute error rate* then avoids this kind of cancellation by removing the polarity (+/-) of the error.

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 human word counts). Annotators for the four ACLEW corpora were proficient in the language spoken in the daylong recording, and transcribed all adult speech in keeping with minCHAT format (e.g., “wanna”, not “want to”; MacWhinney, 2017).

One child in the (otherwise English) SOD corpus was learning French. Given our definition of orthographic words which is not language-specific, we have included this child to increase power, but results without them are nearly identical. See online Supplementary Materials, https://osf.io/zdg6s, for analyses excluding this child. In addition, a total of nine clips from three different WAR children contained some Spanish. Since we are uncertain of how accurate the transcriptions are for Spanish sentences, these clips were removed from consideration altogether.

Human AWC 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 human AWC. The remaining orthographic entries separated by whitespaces were then counted as gold standard target words for LENA® to detect.

The 1- or 2-minute clips sampled for manual annotation were not guaranteed to perfectly align with LENA® segments (i.e. talker onsets and offsets), posing a potential issue for comparing LENA® AWC relative to the human annoted word count. 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 proportionally. That is, 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 analyses, similarly to CVC and CTC.

Table 4

Number of frames, percentage of frames, and number of minutes attributed to each category by the human annotators.

	Frames	Percentage	Minutes
CHI	588,236	7	98
FEM	891,717	10	149
MAL	234,199	3	39
OCH	262,702	3	44
OVL	271,636	3	45
ELE	218,535	3	36
Other	6,112,975	71	1,019

Results

Before starting, we provide some general observations based on the manual human annotations. The "Other" category (meaning no speech, potentially silence but also non-human noise) was extremely common, constituting 71% of the 10 ms frames. In fact, 30% 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 10% of the frames, the child contributed to 7%, and male adult voices, other child voices, and electronic voices were only found in 3% 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 Other (i.e., absence of speech) would be 71% correct. This is of course not desirable, but this fact highlights that systems well adapted to this kind of recording 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 a safe guess is that this speech comes from female adults, who

provide a great majority of the speech, nearly 1.5 times as much as the key child and 2 times more than other children or male adults. In fact, given that speech by male adults and other children is relatively 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 6% of the frames.

LENA® classification accuracy: False alarm, miss, confusion rates. The analysis that yields the best LENA® performance (Table 5, Speakers) focuses on the clean human speaker categories while mapping electronic voices and overlap in the human annotation onto Other, so that the categories considered 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 Other; see Tables 1 and 3). Calculated in this way, LENA®'s *false alarm* rate (i.e., tagging a speech category when there was none) and *confusion* rate (i.e., providing the wrong label) were lowest. Notably, however, the *miss* rate (i.e., the system returns a judgment that no sound label is activated) was double that found with the other analysis alternatives.

In the second-best performing case (Table 5, +Electronic), overlap found in the human annotation is still mapped onto Other but Electronic voices are not, so that the human categories considered were CHI, FEM, MAL, OCH, and ELE; and the LENA® categories considered were CHN, FAN, MAN, CXN, and TVN (with all "far" classes and OLN mapped onto Other).

Finally, performance was worst when included also overlapping regions (Table 5, +Overlap), 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. It is likely that these differences are partially due to OLN and TVN not being defined similarly across the LENA® system and human annotators.

Table 5

*False Alarm Rate (FAR), Miss Rate (MR), Confusion Rate, and total Identification Error Rate (IDER, sum of the medians of the other three categories), as a function of which categories are considered. Speakers indicates that only speaker categories are considers (all others are mapped onto Other); + Electronic that also electronic was scored; + Overlap that electronic and overlap in both human and LENA annotation were also scored. To be maximally informative, we report results in three ways: (1) *weighted by speech*: Overall false alarm, miss, and confusion rates over all clips together, thus giving more weight to clips with more speech; (2) *equal weight per clip*: means across clips, which represent central tendency when giving equal weight to clips with more versus less or no speech; and (3) *accounting for potential outliers*: since means are not robust to outliers, we also report the median across all clips.*

	Overall				Mean				Median			
	FAR	MR	CR	IDER	FAR	MR	CR	IDER	FAR	MR	CR	IDER
Speakers	13	56	11	79	26	39	12	73	6	39	8	73
+ Electronic	44	24	38	106	86	20	36	132	20	12	35	88
+ Overlap	58	22	42	122	126	17	42	172	30	9	41	98

LENA® classification accuracy: Precision and recall. By now, we have established that the best performance emerges when "far" labels such as CHF and OLF are mapped onto Other, as are TVN/ELE and OLN/OVL. False alarm, miss, and confusion rates are informative but 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® users are particularly interested in the key child. The metrics reported thus far do not give more importance to certain classes (such as key child), and they do not give us insight into 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 users 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 classes into a category called "overlap". Please remember that this category is not defined the same way as the LENA® overlap category. For LENA®, overlap was a trained class, and annotators had tagged overlap between two speakers of the same kind (e.g., two female adults) as well as overlap between any of the non-speech classes they were coding (e.g., overlap between noise and TV). We also define overlap as two active classes activated at the same time, but only speech (human or electronic) has been tagged, and can count as overlap in the human annotation.

LENA® classification accuracy: Precision.

We start by explaining how to interpret one cell in Figure 2: Focus on 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 60% of the time. The remaining 40% of frames that LENA® tagged as FAN were actually other categories according to our human coders: 18% were Other (i.e., absence of speech), 10% were in regions of overlap between speakers or between a speaker and an electronic voice, and 12% were confusions with other speaker tags. Inspection of the rest of

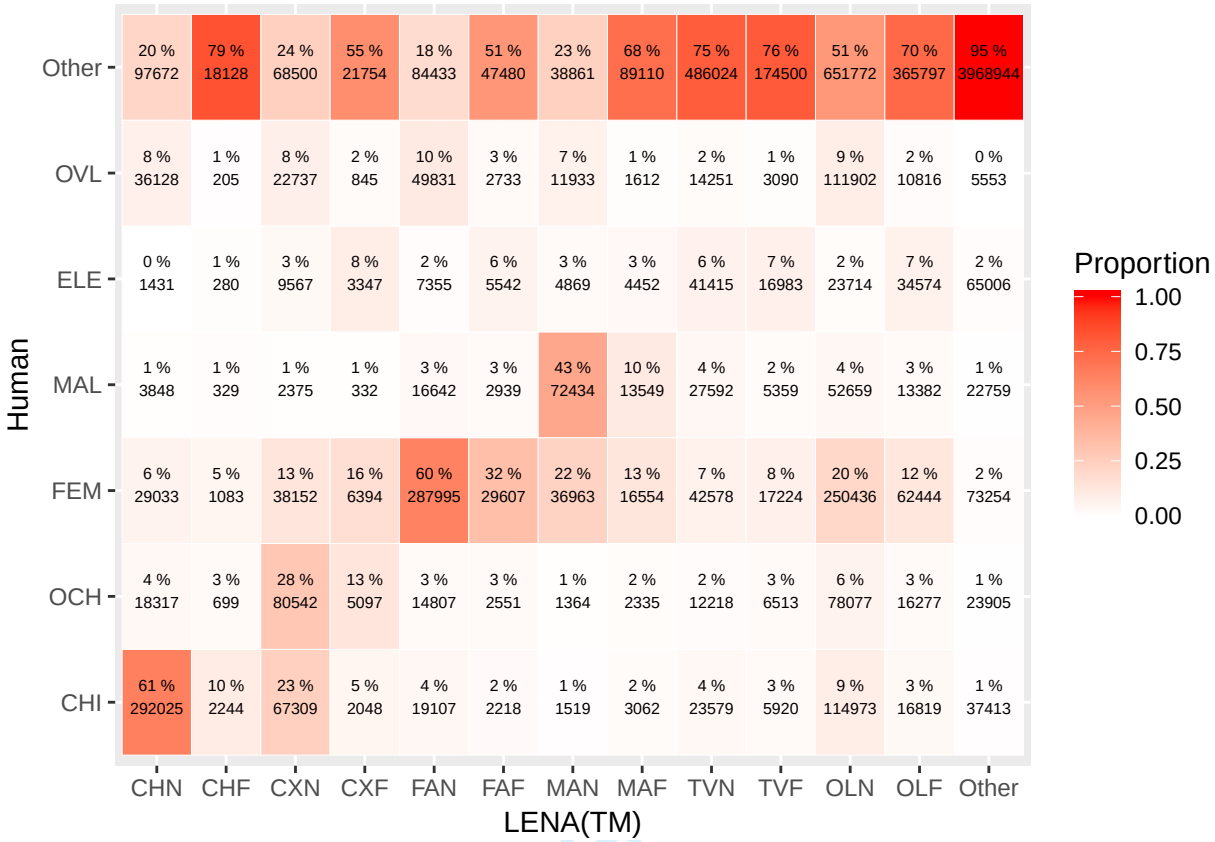

Figure 2. Precision: Confusion matrix between LENA (x axis) and human annotations (y axis). In each cell, the top number indicates the percentage of all frames in that LENA category (column) that are labeled as a given class by the human (row); cells in a given column add up to 100%. The number below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

the confusion matrix shows that FAN and CHN are the LENA® tags with the greatest precision (setting aside the Other class, i.e., lack of speech).

Indeed, precision for CHN is almost identical, at 61%; thus, over 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 Other (that is, silence or more generally lack of speech) according to the human annotator (20%), with the remaining errors being due to confusion with other categories. Aboout 6% of them are actually a female adult; 4% are

another child, and 8% are regions of overlap across speakers, according to our human coders.

Lower precisions are found for MAN (43%) and CXN (28%). The pattern of confusion is somewhat different from the other two categories we looked at, due to greater confusion with the other label within the same age class. That is, 22% of the frames LENA® tagged as being MAN actually corresponded to female adult speech according to the human annotation. It was also not uncommon to find a CXN tag for a frame human listeners identified as a female adult (13%), but even more confusions involved the key child (28%). In a nutshell, this suggests increased 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 other speakers or Other (i.e. silence, absence of speech).

Another observation is that the "far" tags of the speaker categories do tend to more frequently coincide with regions where humans did not detect speech (i.e., Other; 67%) than the "near" tags (35%), and thus it is reasonable to exclude them from consideration for most purposes.

The relatively high proportion of near LENA® tags that correspond to Other (i.e., absence of speech) regions (range 18-75%) 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 likely bias the human annotator to ignore such interstitial silences to a greater extent than if they have no access to the LENA® tags. We inspected how often this padding by the LENA® system occurred and found that it was quite common: About half of the key child's

linguistic and non-linguistic vocalizations tagged in any given clip were shorter than 600 milliseconds long, and thus, if alone, would have been padded by LENA® with silence automatically.

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

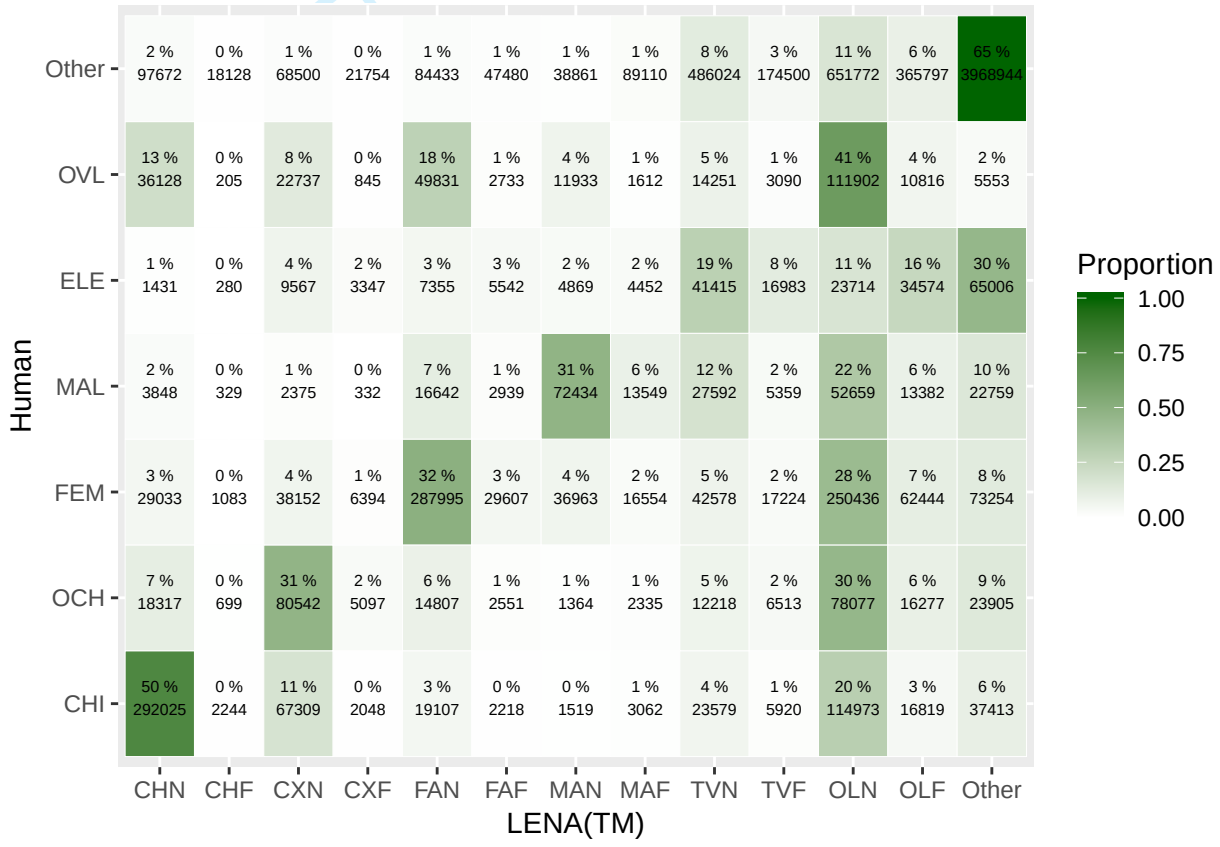

Figure 3. Recall: Confusion matrix between LENA (x axis) and human annotations (y axis). In each cell, the top number indicates the percentage of all frames that a human labeled as a given class (row) which were recovered in a given LENA category (column); cells in a given row add up to 100%. The number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

LENA® classification accuracy: Recall.

Again, we start with an example to facilitate the interpretation of Figure 3. As seen at the intersection of human CHI (last row) and LENA® CHN (first column), the best performance for a talker category for recall is CHN: 50% 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, 11% was captured by the LENA® system's CXN category and 20% 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 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 11% of the key child's speech. Moreover, as we saw above, 28% of the CXN LENA® tags corresponds to the key child, which means that human coders re-coding CXN regions could filter out the 72% that do not, if finding key child speech were a top priority.

Many researchers also use the LENA® as a first pass to capture female adult speech through the FAN label. Only 32% 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 28% of the original female adult speech and, out of the OLN tags, 20% are indeed female adults (meaning that if human annotators are re-coding these regions to find further female adult speech, they would filter out 80% of the segments, on average).

For the remaining two speakers (MAL, OCH), recall averaged 31%, meaning that a third of male adult and other child speech is being captured by LENA®. In fact, most of these speakers' contributions are being tagged by LENA® as OLN (mean across MAN and CXN 26%) or TV (mean across MAN and CXN is 10%), 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 is worth noting that it is 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, 50% were picked up by the LENA® CHN label whereas essentially 0% were picked up 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.

LENA® classification accuracy: Agreement using Cohen's κ.

Given results above suggesting that our coding of electronics may not have coincided with the LENA® system's, and that "far" categories are inaccurate, in this analysis we only consider the following labels for LENA®: CHN, FAN, MAN, and CXN (all others are collapsed into an Other category); and the following labels for human annotators: FEM, MAL, CHI, OCH (all others are collapsed into an Other category). This analysis revealed a Cohen's κ estimated at K(8580000) = 0.44, weighted κ estimated at K(8580000) = 0.46.

Derived counts' accuracy. The accuracy of derived counts (CVC, CTC, AWC) is represented graphically in Figure 4, statistics are provided in Table 6, and error metrics in Table 7.

For CVC, there is a strong association between clip-level counts estimated via the LENA® system and those found in the human annotation, which is not much affected when only clips with some child speech (i.e., excluding 345 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 rather well. The error analyses reveal that, generally speaking, LENA® has a slight tendency to underestimate vocalization counts, particularly when only clips with some child speech are considered. This understimation, however, is not systematic, and cumulating errors using the absolute error rate suggests that the deviation from the actual counts might be quite significant.

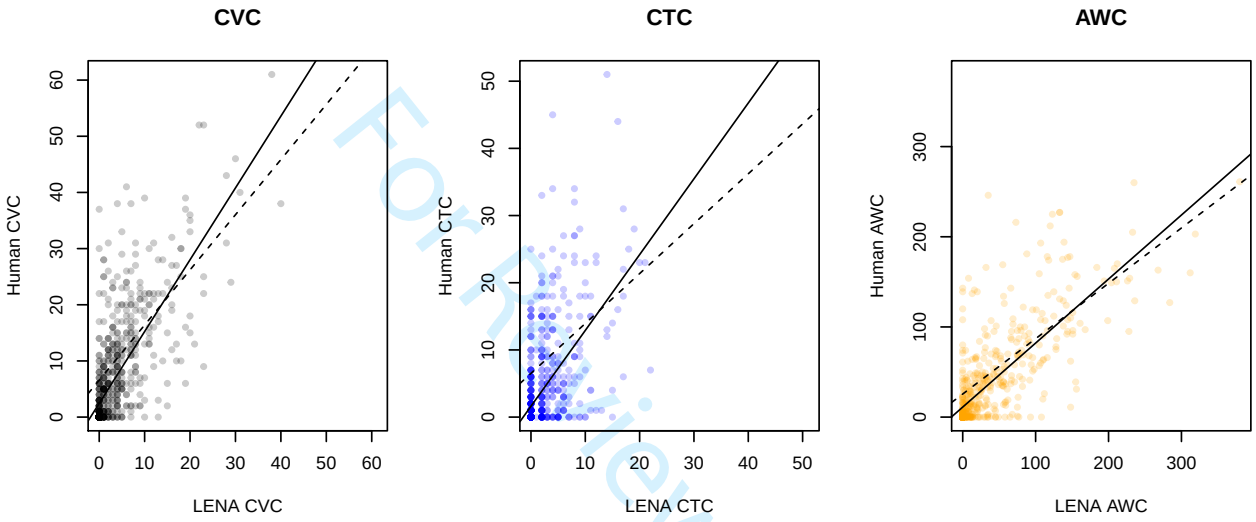

Figure 4. Child Vocalization Counts (CVC), Conversational Turn Counts (CTC), and Adult Word Counts (AWC) according to LENA (x axis) and humans (y axis). Each point represents the counts totaled within a clip. The solid line corresponds to a linear regression fit to data from all clips; the dashed line corresponds to an analysis excluding clips where both the human and LENA® found zero counts. The x and y ranges have been adjusted to be equal regardless of the data distribution.

As for CTC, the association between clip-level LENA® and human CTC was weaker than that found for CVC, particularly when only clips with some child speech (i.e., excluding 208 clips with 0 counts in both LENA® and human annotations) were considered. Inspection of errors and error rates reveals that LENA® tends to underestimate turn counts, which is particularly clear when excluding clips with no turns. As with CVC, the bias varied across clips leading to a substantial cumulative absolute error rate.

Table 6

Number of clips (N) and corresponding Pearson r coeffecient for CVC, CTC, and AWC. 'N all' and 'r all' are computed over all clips. 'N' and 'r' represent non-null clips only (i.e., having some vocalizations, turns, and adult words respectively).

	N all	r all	N	r
CVC	757	0.728	343	0.613
CTC	757	0.567	206	0.351
AWC	589	0.762	301	0.698

The association between clip-level LENA® and human AWC in the four English-spoken corpora was strong, even when only clips with some adult speech (i.e., excluding 303 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 AWC across clips well. Error analyses for AWC reveal a different pattern from before, as the system exhibits a slight tendency to over-estimate AWC. However, this trend was inconsistent, leading to the highest absolute error rate metric among the three derived counts.

Effects of age and differences across corpora. The preceding sections include overall results collapsing across corpora. However, it is 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

Table 7

*Mean (range) for each type of error estimate for CVC, CTC, and AWC. Error estimates are: E (error; NL-NH, where NL means the count according to LENA and NH the count according to the human), E-0 (error excluding clips with a zero count according to human or system analysis), ER (error rate; (NL-NH)/NH*100, in percent of the total), and AER (absolute ER; abs(NL-NH)/NH*100, in percent of the total, with abs meaning that we take the absolute); ER and AER exclude clips where the human count is zero.*

	E	(range)	E-0	(range)	ER %	(range)	AER %	(range)
CVC	-3	(-37,14)	-6	(-35,14)	-39	(-100,650)	74	(0,650)
CTC	-2	(-41,15)	-5	(-41,15)	-29	(-100,1200)	94	(0,1200)
AWC	-1	(-211,157)	-1	(-211,157)	54	(-100,7400)	124	(0,7400)

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 report on key results here; for the full model output and additional analyses, please refer to our online Supplementary Materials (https://osf.io/zdg6s).

Are there differences in false alarm, miss, and confusion rates as a function of corpus and child age?

Figure 5 represents identification error rate as a function of age and corpus for individual children. To test the possible impact of age and corpus statistically, we predicted false alarm, miss, and confusion rates in the analysis with all "Far" categories, TVN/ELE, and OLN/OVL mapped onto Other (which yielded the best results in Section "False alarms, misses, confusion" above.) Our predictors were corpus, child age, and their interaction as fixed effects, and child ID as a random effect. We followed up with a Type III ANOVA to assess significance.

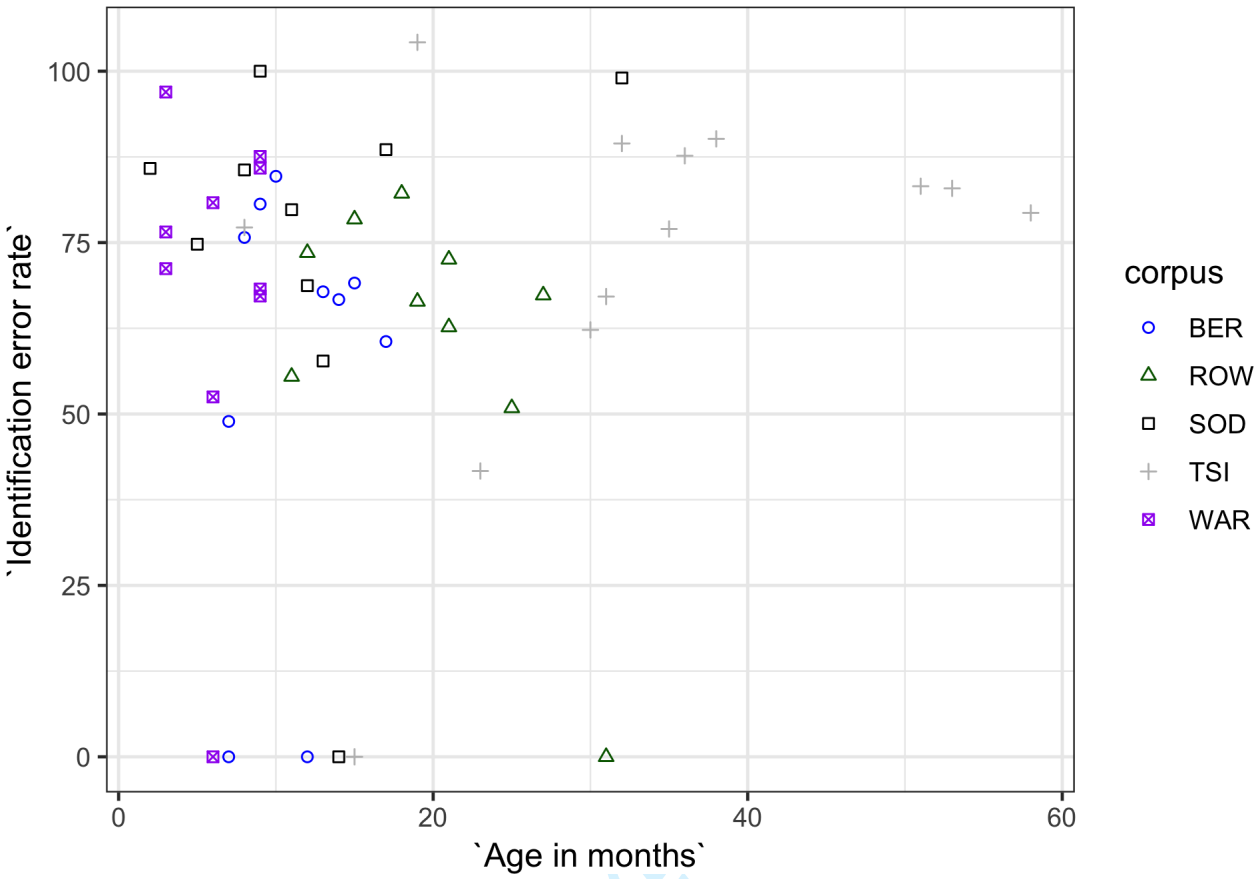

Figure 5. Identification error rate as a function of corpus and child age. Each point represents the median over all clips extracted from the data of one child. Color and shape indicates corpus: BER in blue circles, L05 is green triangles, SOD in black squares, TSI in gray pluses, and WAR in purple crossed squares. A number of the children had a median identification error rate of zero due to the fact that they had many clips in which there was no speech, and LENA had no false alarms, pulling the median to zero.

Corpus, child age, and their interaction were never significant, with the exception of confusion, where the interaction between corpus and age was significant at α=.05. To investigate this effect further, we fit a mixed model predicting confusion rates from child age as fixed and child ID as random effects on each corpus separately. This revealed a main effect of age for SOD only (Chisq (1) = 14.53, p = < .001; all other chi-squares were smaller than 14.53, p > .120).

Table 8

*Results of Type III ANOVAs on false alarms (FA), misses (M), and confusions (C): Chi-square (degrees of freedom), followed by * if the relevant factor is significant (p<.05).*

	FA	M	C
Intercept	0.21 (1)	6.21 (1) *	4.88 (1) *
Corpus	1.98 (4)	3.16 (4)	3.65 (4)
Age	0.06 (1)	0.02 (1)	0.13 (1)
Corpus*Age	0.87 (4)	2.33 (4)	14.39 (4) *

Are there differences in CVC accuracy as a function of corpus and child age?

For CVC, we fit a mixed model where manually-annotated CVC was predicted from LENA® CVC, in interaction with corpus and age, as fixed factors, and child ID as a random effect. Results are summarized in Table 9 (for CVC, CTC, and AWC). Only effects and interactions involving the LENA® predictor are relevant to the present work, and will be discussed here. A Type III ANOVA found a main effect of LENA® CVC, because this was a good predictor of the human CVC.

Are there differences in CTC accuracy as a function of corpus and child age?

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 III ANOVA found a main effect of the LENA® CTC estimates, as well as an interaction between this factor and corpus. We

followed up on this by fitting a model where CTC according to the human was predicted from CTC according to LENA® as fixed and child ID as random factor, for each corpus separately. Inspection of these results (full output available from the Supplementary Materials, https://osf.io/zdg6s) suggests that the interaction emerged because the predictive value of LENA®'s CTC with respect to human counts was stronger for some corpora (Chi-squares for ROW 174.71, BER 107.49, and TSI 99.80) than others (Chi-squares for WAR 57.29, and SOD 30.91; all degrees of freedom are 1, and $p < .001$).

Are there differences in AWC accuracy as a function of corpus and child age?

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 III ANOVA revealed, in addition to a main effect of the LENA® AWC estimates, a three-way and both two-ways interactions involving the LENA® predictor, which were investigated by fitting additional mixed models to each corpus separately. An interaction between LENA® AWC and age was found for BER/WAR as well as SOD, due to a *decreased* predictive value of the LENA AWC with respect to the human AWC for older infants in BER and WAR but an *increase* with age in SOD. However, it should be noted that the association between LENA and human AWC was significant and positive for all four corpora.

Discussion

The aim of the present study was to assess LENA® accuracy across key outcome measures: speaker classification accuracy, Child Vocalization Counts (CVC), Conversational Turn Counts (CTC), and Adult Word Counts (AWC). We did this using an approach that

Table 9

*Results of Type III ANOVAs when predicting human counts (CVC, CTC, AWC) from LENA counts in interaction with age and corpus: Chi-square (degrees of freedom), followed by * if the relevant factor is significant (p<.05).*

	CVC	CTC	AWC
Intercept	5.34 (1) *	0.04 (1)	0 (1)
LENA	8.61 (1) *	19.1 (1) *	46.23 (1) *
Age	0.79 (1)	0.54 (1)	0.6 (1)
Corpus	8.18 (4)	3.41 (4)	1.71 (3)
LENA*Age	0 (1)	2.28 (1)	10.51 (1) *
LENA*Corpus	8.53 (4)	11.9 (4) *	15.7 (3) *
Age*Corpus	5.99 (4)	4.06 (4)	1.59 (3)
LENA*Age*Corpus	6 (4)	4.75 (4)	18 (3) *

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/Other). This permitted a systematic, extensive, and independent evaluation of LENA®'s key automated metrics. We also tested generalizability by analyzing LENA®'s performance 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 by 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 identification error rate), and evaluated the pattern of errors in more detail, by assessing how LENA® and human annotators agreed (precision and recall). The identification error rate was relatively high (global 79, 106, and 122%), mainly due to a high miss rate (missing or excluding speech that was there; 56, 24, and 22%). The false alarm rate (identifying non-speech/silence as speech; 13, 44, and 58%) and confusion rate (identifying voice type; 11, 38, and 42%) were low.

To put these numbers in context, we asked the ACLEW project members to share with us preliminary results of their ongoing inter-rater reliability study. This study covers six corpora, including the four ACLEW corpora used here. For their reliability analyses, they considered the "gold" to be the original complete annotations, and the "system" the reliability annotations, which were done later and in only a subset of the corpus (one minute per day-long recording, for a total of 60 minutes across their six corpora). While we cannot report on these results in full because their publication is intended elsewhere, we can state the following overall observations. Among two human annotators, the ACLEW team reported an identification error rate of 56% (due to 20% false alarms, 19% miss rates, and 17% of confusion); for the four databases included here, the average identification error rate was 47%. This is considerably lower than the identification error rates reported for LENA® here (best case scenario yielding a identification error rate of 79, 106, and 122%), mainly due to a much lower miss rates, whereas both false alarm rates and confusion rates are higher across the two human coders. Inspection of false alarms and misses suggests the disagreement across humans emerges when there is background speech, that one coder may pick up on and not the other.[2]

[2]Taking all categories together, Cohen's κ agreement was .64 (weighted κ .65) for the ACLEW inter-rater reliability coding on all six ACLEW datasets, which is higher than the best case scenario for LENA (.46).

Another question is how LENA® fares compared to other automatic systems. Our thorough review of the literature revealed that no previous report is comparable: Most often, the data used is considerably different (and overall easier; e.g., recorded in formal settings, with a small number of speakers, who produce long vocalizations); moreover, previous research tends to overestimate performance by using lax evaluation criteria (e.g., allowing errors in a "collar" around each vocalization). The most comparable data point comes from the DIHARD Challenge (Ryant et al., 2019). DIHARD employed data from a range of domains, including daylong recordings; in fact, they used a different selection of data from the BER corpus used here. The subset of BER used for DIHARD is likely to lead to lower error rates because they selected only files that contained some speech; by excluding files with little to no speech, they prevent the appearance of very high diarization error rates (which emerge when the numerator, i.e. the amount of speech, is very small). Thus, the DIHARD reanalyses are likely to overestimate the systems' performance in terms of data selection. Their evaluation, however, was as strict as ours, with no leeway or collar. Diarization error rates for the BER subset by systems submitted to DIHARD 2019 varied between 48% and 121%, with a median around 70%. Thus, LENA® is competitive with respect to state-of-the-art systems, although some of them do score considerably better.[3]

Returning to the LENA® system results, the overall error rate can be fruitfully interpreted by considering performance on individual speaker tags. 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 (60% of frames tagged by LENA® as FAN were coded as female adult by the human coders), and the target child (61% of frames tagged by LENA® as CHN were correct). However, the system performed substantially worse with

[3]DIHARD uses diarization error rate on individual speakers' identities, rather than identification error rates on speaker types as we do here. There is no mathematical procedure to derive one from the other, except in the case when there is one speaker per speaker type, in which case diarization error rate is most likely identical to identification error rate.

other talker types (e.g. 43% and 28% for MAN and CXN, respectively); that is, less than a half of the frames that LENA® tagged as being speech spoken by these speakers actually correspond to them.

To get a sense of how these results compare to multiple human coders, we also asked about precision and recall in the reliability data from the ACLEW team. Across all six corpora, precision for key child was the highest, at 80%; for the other speakers it was: 72% female adult, 72% male adult, and 65% other child. Precision is higher and more similar across speaker types in the ACLEW reliability data than in our LENA®-human comparison here.

In terms of recall (how accurately LENA® captured the human annotations), performance for the key child's vocalizations was moderately robust: 50% of the frames humans attributed to the key child were captured by LENA® under the CHN tag. However, recall was poorer for the other three talker types, at around 31-32%. As for recall in the ACLEW reliability data, the key child score was 79%; for the other speakers it was: 71% female adult, 63% male adult, and 55% other child. Thus, although we see lower recalls for male adults and other children in both, the overall level of recall is much higher across two human coders than between LENA® and human, mainly due to LENA®'s tendency to miss speech more than humans do. This is, however, sensible for a system aimed at analyzing day-long recordings, which contain long stretches of silence.

Our second set of analyses tested the accuracy of three of the aggregated counts automatically provided by LENA®, namely CVC, CTC and AWC. We found relatively 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 estimates. For overall error estimates (i.e., how far was the LENA® count from the human annotators'), the means

across clips for CVC, CTC, and AWC was an encouraging -3.32,-1.85, and -1.04, respectively. These low errors were not solely due to many clips lacking vocalizations, turns, or adult words altogether, because when we exclude such clips we still get what seem to be low errors: means were -6.46,-5.08, and -0.67 for CVC, CTC, and AWC respectively.

We also put these error patterns in context by taking into account how large the counts were to begin with. Such error rates, however, are only defined for files which, according to the human, contain at least one unit (otherwise, we divide an error of a certain size by zero, which is undefined). We find error rates suggesting that LENA® counts are off by between a third or a half of the original counts. Inspection of the sign in these rates indicates that, 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. In addition, the absolute error rate analysis, which prevents under- and over-estimations from cancelling each other out, reveals rather considerable errors.

Finally, we also inspected the extent to which LENA® performance was affected by dialect, language, and child age in a final set of analyses. We would like to be tentative about the interpretation of these results, because we only have about 10 children, often varying widely in age, in each corpus, with some mismatch in age range across corpora (see Table 2). This means that we did not have a great deal of power to capture true differences across corpora and that we may have some spurious effects or interactions due to chance differences.

With all these caveats in mind, 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. This is not what we found. For instance, we found an interaction between corpus and age for the confusion rate, due to an increase in confusion

rates for older infants within the SOD corpus but not in any of the others, a result that we have no ready explanation for, and which may be a spurious result given the sample size (10 recordings from 9 children, in this corpus). Similarly, LENA® counts predicted human counts in the CVC, CTC, and AWC analyses, and although we did observe some interactions, none of them were easy to interpret and none explained away the predictive value. As just mentioned, we are cautious when interpretating these results, and invite further work on bigger samples (more data per child, more children per corpus) to ensure sufficient power and precision.

In general, whether LENA® results are "good enough" depends largely on the goals of each particular study. For example, we can describe precision rates of 60% (i.e., 60% of frames tagged by LENA® as FAN were coded as female adult by human coders) and 61% (i.e., 61% of frames tagged as target child were also tagged as such by human coders) as being reasonably good, because they are much higher than the system's precision rates for other speakers (MAN 43%, CXN 28%). Although they are lower than what may be found across two human raters, some additional level of error may be expected in an automatic system. Notice saliently that, despite having been created over a decade ago, the global performance of LENA® was competitive when compared to state of the art diarization systems. 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 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 "female adult" 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: the 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 fairly good at capturing speech by the key child as such, but recall was lower for the other talker categories.

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 in 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 23% 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® identification 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 a small sample from a handful of corpora, it seems that the system does not perform a great deal worse for children who do not correspond to the LENA® training set. Moreover, *correlations* between human and LENA® clip-level counts were high, suggesting that the software accurately captures differences in counts across clips (even when *absolute error rates* were also high). Except for CTC, these correlations remained quite high even when clips with counts equal to zero were removed from consideration, suggesting that LENA® captures gradience in vocalization and adult word counts.

However, our finding that LENA® generally underestimates the quantity of child vocalizations and child-adult turns deserves further consideration. 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). If underestimates are 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.

How to test the reliability of the automated output provided by LENA®. 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 ideal 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 corresponding LENA® output file.

First, we recommend a literature search [starting from A. Cristia et al. (2019)'s a systematic review], to determine whether there exists reliability data for a similar sample. If no reliability studies exist, draw 10 x 2 minutes randomly from 10 children. This is about 3h20min of data, which takes roughly 90h 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 code we provide in supplementary materials (https://github.com/jsalt-coml/lena_eval). This will allow researchers to extract the classification accuracy measures used here (false alarm rate, miss rate, confusion rate and the derived identification error rate), as well as CVC, CTC, and AWC comparing LENA® and human annotations. We note re-using our code 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.

One issue that may arise is whether data should be sampled differently to, for example,

make sure every class is represented the same amount of time and/or a minimum of time. Our understanding is that class imbalance and data scarceness is an important issue for training, and directly affects algorithm accuracy (this is a general problem, but to cite just one example on GMMs, Garcia-Moral, Solera-Urena, Pelaez-Moreno, & Diaz-de-Maria, 2011). However, it does not pose the same kind of problem for evaluation. That is, if there are no samples of a given category, then accuracy cannot be evaluated; if there are only a few, then it is possible that these are special in some way and accuracy estimates may not generalize well to others. Thus, it would indeed be desirable to have enough samples of a given label to reduce the impact of each individual instance, in case they are outliers. That said, almost any strategy that attempts to boost the frequency of specific categories risks worsening non-generalizability concerns. For instance, if one were to over-sample regions tagged by LENA® as MAN in the hopes of having more male samples, one may only be capturing certain types of male speech or acoustic properties. To take this example further, notice that male speech is our smallest category, representing 3% of the data. Since we sampled randomly or periodically, this represents the prevalence of male speech and the samples that are included are unlikely to be acoustically biased.

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 difference 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), in a sample of data drawn across different dialects, languages, ages, and socio-cultural 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 are

not deidentified); these are stored in private repositories that do not have a persistent identifier. The annotation data were used to generate statistics at the clip level, which are the input to analyses presented here. Both the clip level statistics and analyses in this manuscript are publicly available from https://osf.io/zdg6s. None of these analyses were pre-registered. Additional computer code for other levels of analysis is available from https://github.com/jsalt-coml/lena_eval.

References

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. *arXiv Preprint arXiv:1402.1894*.

Bergelson, E. (2016). Bergelson Seedlings Homebank corpus. https://doi.org/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. https://doi.org/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.

https://doi.org/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. (2019). Accuracy of the Language Environment Analysis System: A systematic review. Retrieved from https://osf.io/fhs57

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

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.

Garcia-Moral, A. I., Solera-Urena, R., Pelaez-Moreno, C., & Diaz-de-Maria, F. (2011). Data Balancing for Efficient Training of Hybrid ANN/HMM Automatic Speech Recognition Systems. *EEE Transactions on Audio, Speech, and Language Processing*, *19*(3), 468–481. https://doi.org/10.1109/TASL.2010.2050513

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. https://doi.org/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. https://doi.org/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. https://doi.org/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.

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. https://doi.org/10.1073/pnas.1003882107

Orena, A. J. (2019). Growing up bilingual: Examining the language input and word segmentation abilities of bilingual infants. PsyArXiv. https://doi.org/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. https://doi.org/10.17605/OSF.IO/KAU5F

RStudio Team. (2019). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

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. (2019). Daylong audio recordings of young children in a forager-farmer society show low levels of verbal input with minimal age-related change. *Manuscript in Progress.*

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. (2019). The ACLEW Random Sampling corpus. *Manuscript in Progress.*

Team, R. C., & others. (2013). R: A language and environment for statistical computing.

Vienna, Austria.

VanDam, M., & De Palma, P. (2018). A modular, extensible approach to massive ecologically valid behavioral data. *Behavior Research Methods*. https://doi.org/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. https://doi.org/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.