Describing vocalizations in young children: A big data approach through citizen science annotation

Chiara Semenzin[1], Lisa Hamrick[2], Amanda Seidl[2], Bridgette Kelleher[2], & Alejandrina Cristia[1]

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

[2] Purdue University

Author Note

AC acknowledges Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. LH acknowledges the National Institute of Deafness and Other Communication Disorders (F31DC018219). BK acknowledges the National Institute of Mental Health (K23MH111955) and the Kinley Trust. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. The funders had no impact on this study. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work and have no conflict of interests to disclose.

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

Abstract

Purpose: Recording young children's vocalizations through wearables is a promising method. However, accurately and rapidly annotating these files remains challenging. Online crowdsourcing with the collaboration of citizen scientists could be a feasible solution. In this paper, we assess the extent to which citizen scientists' annotations align with those gathered in the lab for recordings collected from young children.

Method: Segments identified by LENA^{TM} as produced by the key child were extracted from one daylong recording for each of 20 participants: 10 low-risk control children and 10 children diagnosed with Angelman syndrome, a neurogenetic syndrome characterized by severe language impairments. Speech samples were annotated by trained annotators in the laboratory as well as by citizen scientists on Zooniverse. All annotators assigned one of five labels to each sample: Canonical, Non-Canonical, Crying, Laughing, and Junk. This allowed the derivation of two child-level vocalization metrics: the Linguistic Proportion, and the Canonical Proportion.

Results: At the segment level, Zooniverse classifications had moderate precision and recall. More importantly, the Linguistic Proportion and the Canonical Proportion derived from Zooniverse annotations were highly correlated with those derived from laboratory annotations.

Conclusion: Annotations obtained through a citizen science platform can help us overcome challenges posed by the process of annotating daylong speech recordings. Particularly when used in composites or derived metrics, such annotations can be used to investigate early markers of language delays in non-typically developing children.

Describing vocalizations in young children: A big data approach through citizen science annotation

Since early language delays can adversely affect children's literacy, behavior, social interaction, and scholastic achievement extending well into adulthood, early interventions have been described as a better societal investment than later ones (Heckman, 2006). Some research suggests that spontaneous behavior, captured for instance via home videos, could provide important indices to development (Belardi et al., 2017; Overby, Belardi, & Schreiber, 2020). Advancements in the field of wearable technologies, such as LENA™ recorders, have opened new avenues to both early detection of speech pathologies and research in language development more generally (Oller et al., 2010; Rankine et al., 2017; VanDam & Yoshinaga-Itano, 2019). Wearable recorders allow data collection to happen in the child's natural environment, and at a large scale, which may be particularly helpful for children whose speech is not easily elicited. Although such long-form recordings are increasingly common (Ganek & Eriks-Brophy, 2018), challenges remain with respect to how these data are handled, annotated, and analyzed (Casillas & Cristia, 2019). The present work reports on the validity of labels and child-level descriptors of children's vocalizations gathered from citizen scientists, in a comparison to expert lab annotators.

In the rest of the Introduction, we first briefly summarize two metrics that can be used to describe individual children's vocalizations in the context of long-form recordings. We then introduce crowdsourcing in general, and crowdsourcing by citizen scientists in particular, as a potential avenue for more rapidly deriving these key metrics from vocalization recordings.

Describing children's vocalizations in long-form recordings. A large body of research has investigated both fine-grained and coarse descriptions of children's vocalizations as a function of age and diagnosis, and it is beyond the scope of this paper to provide a full summary of this work [but see **XX ADD REFERENCE** for a recent summary].

Moreover, the way in which vocal development has been studied in the past may not easily translate to long-form recordings. As mentioned earlier, long-form recordings have several advantages, including capturing the child's vocal patterns in their natural environment and being able to accumulate a great amount of data easily, which is particularly useful for diagnoses and pathologies characterized by low levels of vocal production. However, this also means that the audiorecording is harder to process than an audiorecording gathered in more manicured and stable conditions: In a typical daylong recording, the child will sometimes be in a quiet room with just her primary caregiver, but later in a noisy supermarket or having dinner with family and friends. In fact, human annotation of such audio is estimated to take about 30 times the audio length, with these estimates being greater the more precise one wants or needs to be (Casillas et al., 2017). It is for this reason that most users of long-form recordings have turned to automatized software for at least a first-pass analysis.

Automated first-pass analyses. The most commonly used software was created by the LENA™ Foundation, and it returns a diarization of the audio signal split into key talkers, including the child wearing the device, as well as a split of these audio attributed to the child into three categories: crying, other fixed and vegetative signals, and speech vocalizations. These subcategories have not been widely validated (Cristia et al., 2020). Moreover, LENA™ lacks an important distinction between more and less advanced vocalizations.

Other automatized algorithms have been developed in the last two years, saliently ones attempting to classify child vocalizations into crying, laughing, canonical, and non-canonical (Schuller et al., 2019). The split of speech vocalizations into canonical and non-canonical, and the overt inclusion of laughing, allow a somewhat finer-grained picture of children's vocal activity than what the LENA™ algorithm currently offers. Another advantage of this new line of research is that there exists a challenge where performance has been benchmarked, so unlike LENA™'s algorithms, we know which algorithms perform better than others. Indeed, the ComParE 2019 BabySounds sub-challenge established a state-of-the-art baseline with a test-set unweighted average recall of around 55%. By re-using the same test set, researchers can develop new algorithms and prove that they can outperform this baseline (or any subsequent best-performing algorithm benchmarked in the same test set).

Table 1 provides an idea of the accuracy across different methods. Specifically, we calculated accuracy of the LENA™'s subtypes[1] based on the laboratory annotations used in the present study; and we also show the results of ComParE 2019 BabySounds sub-challenge, although note it is not based on the same data as the other two columns. Table 1 shows that LENA™ has a good recall for speech-like vocalizations and crying but not laughing, and by design it does not distinguish between canonical and non-canonical vocalizations. The ComParE 2019 BabySounds sub-challenge's baseline model achieved a more moderate recall on crying than LENA, but clearly outperforms the LENA™ algorithm in the Laughing category, with the additional advantage of distinguishing canonical and non-canonical – although note that performance in non-canonical is barely above a chance level of 20% (based on 5 categories, assuming equal probabilities). Notice that the baseline model also included a category called "Junk" for sections that turned out not to be the child's vocalizations after all, but LENA™ here has a low recall because only sections that LENA™ classified as being child vocalizations are considered in the present study, and thus this class's category cannot but be low-performing for LENA™. That is, since we selected only segments attributed to the key child, any segment that turns out not to be a child vocalization is a LENA™ error. We return to the comparison between LENA™, the ComParE 2019 BabySounds sub-challenge baseline, and our Zooniverse-based method in the Discussion, but for now, we simply point out that there are new algorithms being developed, but their performance –although competitive when compared to LENA™– is underwhelming.

[1]If a child segment contained speech, then it counted towards canonical/non-canonical; else, if it contained crying it counted towards crying; else, if it contained some fixed signals it counted towards laughing; a small proportion were left that did not have any of the three and were considered as "Junk" or not categorized.

—– INSERT TABLE 1 HERE —–

Derived metrics. One possibility that has only recently begun to be explored is the use of derived metrics. So instead of evaluating algorithms and other annotation procedures on their accuracy at individual labels, one can derive a metric that more closely relates to vocalization development.

For example, most work using LENA™ software reports on children's vocalization counts – this is a derived metric because it is not simply a description of sections of the audio, but instead it integrates over a certain period of time. Although this has been sometimes criticized as being more about quantity than quality (McDaniel, Yoder, Estes, & Rogers, 2020), this is a promising metric of individuals' vocalization development because it shows correlations with age (which is a proxy of development), and it can be extracted quite accurately with LENA™ (Cristia et al., 2020). Additionally, the child vocalization count metric has been found to be concurrently and predictively correlated with an effect size $r \sim .3$ with standardized language scores in a meta-analysis (Wang, Williams, Dilley, & Houston, 2020). Thus despite representing a composite of skills, vocalization count may provide a useful estimate of vocalization development.

Here, we focus on two alternative composite metrics. One metric that can be derived once children's vocalizations are split into crying, laughing, canonical, and non-canonical is the Linguistic Proportion: the proportion of vocalizations that are linguistic (canonical and non-canonical) out of all vocalizations. To our knowledge, this metric has not been extensively explored previously, but it is likely that this proportion increases with age.

Yet another metric that could be extracted to estimate relative linguistic complexity is the Canonical Proportion: the proportion of vocalizations that contain a canonical transition or syllable out of all linguistic (canonical and non-canonical) vocalizations. One recent study has established that the Canonical Proportion extracted from daylong recordings is significantly correlated with age in a multicultural and multilingual sample (Cychosz et al., 2019). Importantly, this association was established in a sample of children going up to 3 years of age, and thus well beyond the babbling period, and into first words and word combinations.

There is a wider research base documenting the potential importance of the Canonical Proportion, and related metrics, although this work focuses on babbling and thus typically on infants under one year of age. A critical milestone involves the increasingly common production of canonical syllables, consonant-vowel or vowel-consonant sequences that resemble those found in adult speech (Oller, Eilers, Neal, & Cobo-Lewis, 1998). Given its adult-like consonant-vowel or vowel-consonant structure, canonical babble is considered to be a starting point on the path to recognizable speech. Canonical syllables show a higher complexity given the smooth articulatory transition between a consonant and a vowel (or vice versa), when compared to more primitive sounds such as squeals, or isolated vowels. Some work suggests that Canonical Proportions above .15 are expected by about 10 months of age in typical development (Oller, 2000). In addition, the proportion of vocalizations containing a canonical syllable has been found to be more predictive of individual development than sheer vocalization counts in a sample of children diagnosed with an autism spectrum disorder (McDaniel et al., 2020).

Crowdsourcing: A potential solution for annotation. In sum, there is a growing literature attempting to use data from daylong recordings to describe young children's vocalizations, but there are two outstanding challenges. The first pertains to how the data are annotated, with human annotation being costly. The second relates to how descriptors of vocal development are generated, i.e., how annotations of audio sections are integrated into vocal development metrics.

Crowdsourcing refers to the process whereby a task is solved by a crowd, rather than an individual. A number of fields, particularly in the data-driven sciences, have already engaged in the collection of data (including annotations) through crowdsourcing, thanks to its low cost and ecological value (Crump, McDonnell, & Gureckis, 2013; Sescleifer, Francoisse, & Lin, 2018). For example, a systematic review on crowd-sourced ratings of speech found that "lay ratings are highly concordant with expert opinion, validating crowdsourcing as a reliable methodology"; across studies, crowdsourced and expert listener classifications yielded a mean correlation coefficient of .81 (Sescleifer et al., 2018). On the other hand, the systematic review returned only 8 studies (of which only four were published in peer-reviewed journals), suggesting that there is considerable need for further research on this topic.

Mechanical Turk (MTurk) MTurk is an online labor market created by Amazon to assist "requesters" in hiring and paying "workers" for the completion of computerized tasks. Although it is a leading crowdsourcing service, and some evidence suggests MTurkers' annotations can be quite reliable (Berinsky, Huber, & Lenz, 2012), some question marks are raised as some "workers" turn out to be bots, or are poorly motivated (and potentially exploited) humans.

Citizen scientists to the rescue. A promising crowd-sourcing alternative has arisen in recent years: citizen science, a research technique that engages the public in the collection of scientific information. As citizen scientists do not receive compensation, this alternative to platforms such as MTurk can overcome the limitations posed by potential exploitation of workforce and/or the use of bots. Volunteers are entirely motivated by the desire to contribute to research advancements as well as the pleasure they derive from the task itself.

One of the most successful platforms hosting citizen science projects is Zooniverse (zooniverse.org; Borne & Zooniverse Team, 2011). The website hosts a multitude of interdisciplinary projects that have allowed the public to take part in cutting-edge scientific research, from marine biology to papyrology. Zooniverse has proven extremely useful in those fields where the complexity of the data collected is too high to be automatically interpreted using computer algorithms. At the same time, the tasks that human volunteers are asked to complete are sufficiently simple that citizens can carry them out without a background in science or any extensive training.

Citizen science may be particularly helpful when analyzing infants' data from wearables, which remain challenging to annotate as mentioned above. There is one previous study that attempted this approach. Cychosz et al. (2019) drew child vocalization data from a diverse set of corpora centered on children learning one of four languages: English, Tseltal, Tsimane', and Yélî. For the English and Tsimane' corpora, vocalizations were automatically identified using LENA™, whereas the other two were extracted through manual segmentation. Segments were then split into maximally 500-ms long clips, and presented to annotators through the citizen science iHearUPlay platform, with the aim that each clip received three classifications into Crying, Laughing, Canonical, Non-Canonical, or Junk (with the latter tag used for clips that did not contain a child voice). They then derived an implementation of the Canonical Proportion, as the proportion of clips receiving a majority judgment of canonical out of the clips receiving a majority judgment of canonical or non-canonical. They found that this Canonical Proportion increased with infant age in an analysis collapsing across corpora, which provides a first proof of principle that citizen scientists' annotations could be an appropriate solution to the problem of annotating child vocalizations from wearables. The authors also studied how their Canonical Proportion related to age within each corpus that had a sufficient number of children varying in age. They found that in one English corpus bearing on very young infants, the proportion of clips assigned to the "Junk" category was very high. Additionally, they found that the correlation between Canonical Proportion and age was much weaker in two corpora where vocalizations were segmented with LENA™ than through manual segmentation. In sum, these analyses within corpora reveal that data may not be uniformly useful, with potential variation across corpora, children, and methods.

The present work. Our work seeks to broaden the already promising results that have emerged from previous work summarized above. We hope that this methodology will open the road to larger scale analyses of children's vocalizations as captured by wearables. In addition, we went beyond Cychosz et al. (2019)"s study in two important ways. First, we relied on the largest and best established citizen science platform: Zooniverse hosts more than 1.6 million users from diverse walks of life, and it offers a completely automatized API system to more easily scale tasks in a transparent and cumulative science fashion. In contrast, Cychosz et al. (2019) actively recruited people to provide annotations, stating that '[a]nnotators included language and speech researchers, undergraduate students, research assistants, and other interested parties, totaling 136 unique annotators'. A process that draws from research assistants relies on expert resources, whose time is expensive. Moreover, the people these researchers recruit tend to be more accustomed to child vocalization data, which may improve their annotation performance, and thus provide an overestimate of the quality of the annotations that can be realistically done within a truly citizen science framework. Those authors also had to rely on manual extraction of the resulting data from the iHearUplay team, which added to the administrative load. Second, a core goal of the present work was to determine how citizen scientists" annotations fare compared to the current gold standard, laboratory annotations. Cychosz et al. (2019) did not have a gold standard for their data.

Thus, the present study aimed to contribute a key piece of evidence missing in this discussion: To what extent do laboratory and citizen science annotations agree when describing young children's vocalizations? We examined the extent to which such classifications agree, by quantifying the correspondence across these two modes of annotations at the level of individual segments, and at the level of individual children.

At the segment level, we checked the extent to which laboratory annotations made by experts agreed with judgments made by citizen scientists. We used confusion matrices to describe annotation convergence and divergence patterns, and overall accuracy, kappa, and Gwet's AC1 coefficient as statistical descriptors. Given previous results suggesting that performing this classification based on local acoustic cues is challenging (Schuller et al., 2019; Seidl, Warlaumont, & Cristia, 2019), we expected agreement to be only moderate at this level.

When considering correspondence at the level of individual children, we derived two metrics: the Linguistic Proportion (Canonical+Non-Canonical)/(all non-Junk judgments); and the Canonical Proportion (Canonical)/(all linguistic judgments).

It was important for generalizability purposes to examine data from children with variable biological, mental age, and vocal maturity levels. We therefore included a group of low-risk control infants, who are close to the population most commonly sampled in studies using LENAˆTM (see meta-analyses in Wang et al., 2020; Cristia et al., 2020); as well as children who had been diagnosed with Angelman syndrome and whose ages spanned a wide age range. Some previous work had shown that vocal development proceeded differently when compared to a normative sample (Hamrick et al., 2019; Hamrick & Tonnsen, 2019). Generally, studies on vocal development among children with genetic disorders reveal lower volubility and lower prevalence of canonical syllables (Belardi et al., 2017; Rankine et al., 2017; see also **???** for a review of retrospective video analysis of children diagnosed with Autism Spectrum Disorder, Rett Syndrome, or Fragile X). Low volubility is one of the contexts in which daylong recordings may be particularly advantageous over alternative recording methods (see Rankine et al., 2017 for discussion), since it is challenging to measure vocal development accurately based on shorter recordings (as the child may not vocalize at all in a 1-2h recording). Additionally, the ideal method will be sensitive not only to the large changes that occur over normative development, but also to the potentially smaller individual differences found among children. Such a level of sensitivity is important if these methods are ever to be used to detect potential effects of intervention.

Returning to our overall predictions with these goals in mind, we were particularly interested in describing potential differences in accuracy of derived metrics as a function of our two participant groups. Previous work suggests that children with Angelman syndrome show decreases in Canonical Proportion with age, whereas in typically-developing infants, conversely, the Canonical Proportion was expected to increase with age (Hamrick et al., 2019; Hamrick & Tonnsen, 2019). Although we could not rely on previous work to make predictions regarding the Linguistic Proportion, we reasoned that we should observe an increase in the Linguistic Proportion for both populations. More specific to our research aims, we checked the degree to which laboratory and citizen science annotations converged at the child level by using correlations across the two.

Methods

All analyses and key data are available for reproduction on OSF (https://osf.io/57yha/). Computer code used to collect the Zooniverse judgments is available from GitHub (Semenzin & Cristia, 2020). This manuscript is reproducible thanks to the use of RMd (Baumer & Udwin, 2015) and Papaja (Aust & Barth, 2017) on R (Team & others, 2013).

Participants. The full data set includes data from 20 children: 10 English-speaking children (6 males, 4 females; age range 11-53 months, mean=41.5 months) diagnosed with Angelman syndrome and 10 low-risk control children (6 males, 4 females; age range 4-18 months, mean=11.7 months). As children diagnosed with Angelman syndrome typically have severe cognitive and language delays and most do not produce more than 1-2 words consistently, we compare their linguistic production data against that of younger children with theoretically similar language profiles. This study was approved by the local institutional review board under the project name Neurodevelopmental Natural History Study, Purdue University, IRB-1811021381.

Equipment and data collection procedure. Recordings were obtained with the Language ENvironment Analysis (LENA™) Digital Language Processor, a lightweight recorder ($<$ 60 g, 5.5 cm x 8.5 cm x 1.5 cm) designed to be worn inside a breast-pocket of purpose-made clothing. In our data, every child contributed at least one daylong recording, with one child contributing two recordings (duration 11.58-16.00 hours, mean 15.05 hours).

Data preprocessing (LENA™). The first step we took before beginning the annotation process was the identification of children's vocalizations within day-long recordings. We did this automatically, using the LENA™ proprietary software (Xu et al., 2008). Specifically, the LENA™ software attempts to assign segments of speech either to the key child (the one wearing the recorder), or to one of 15 other categories in the child's environment (e.g., Female-Adult, TV). Key child labels have a precision of about 60% and a recall of about 50% (Cristia, Lavechin, et al., 2020). These stretches of audio assigned to the key child can contain cries, vegetative sounds, and linguistic vocalizations, as well as interstitial pauses. We will refer to them as *segments*, to highlight the fact that they are stretches of audio segmented out as belonging to the key child, but they may not be individual vocalizations.

Annotation by lab experts. Annotation was carried out at the segment level during lab coding. A subset of segments LENA™ identified as the key child were selected for human annotation. Segments were selected from 30 five-minute sections of the child's recording: 10 sections during which the child was vocalizing at their highest rate, and 20 sections drawn randomly from the remainder of the recording. To ensure generalizability of our data, we included in this subset a number of segments of audio coming from both periods of high child volubility, as well as randomly selected periods of the day.

Annotations were carried out by 20 undergraduate students working in either Speech, Language, and Hearing Sciences or Psychological Science laboratories. Before annotation, coders completed an ethical conduct of research course and received HIPAA compliance training. After this, they completed a brief training on vocal maturity that involved 1) reading relevant literature, 2) reviewing examples of various annotation categories and 3) submitting answers to a brief quiz to test their annotation accuracy. Coders then proceeded to the classification, where they made decisions on whether a segment is a cry, laugh, non-canonical syllable(s), canonical syllable(s), or a word. Coders were instructed to classify each segment based on the highest level of vocal maturity it contains (e.g., a segment with both canonical and non-canonical syllables would be classified as canonical; a segment with both laughing and non-canonical syllables would be classified as non-canonical). For canonical, non-canonical, and word segments, coders indicated how many syllables of each type of vocal maturity they hear (e.g., "ah ma ba" would be 1 non-canonical and 2 canonical syllables). For the purposes of the present study, we considered a "word" judgment to be equivalent to a "canonical" judgment. This may be false sometimes (e.g., "oh" may be coded as a real word by laboratory experts, although it is a non-canonical syllable), but we could not have a "word" judgment in our Zooniverse annotations because that requires context (which compromises privacy – see next section).

Coders are also given a "Don't Mark" option, which they are instructed to use if the segment does not sound like a segment made by the target child, or if there is any overlapping speech or other noise in the background which could affect acoustic analyses. Notice, therefore, that this category is broader than the "Junk" category used by Zooniverse annotators.

This process resulted in high levels of inter-rater agreement: Two or more of the three total coders who annotated that segment agreed on 97.22% of the lab-annotated segments.

Annotation on Zooniverse. A substantial concern that emerges when considering citizen science annotation for spontaneous speech data is the risk of a privacy breach: even short clips can contain personal information that can expose the identity of the speaker, such as addresses or names, or contain sensitive and private information. Fortunately, Seidl et al.

(2019) determined that when segments are divided into shorter clips (400-600ms), human annotators with little training can code our categories of interest (Canonical, Laughing, Crying, etc.) with a classification quality comparable to the one carried out on full segments. We therefore used very short clips (500ms) as these are unlikely to contain more than two syllables, and thus prevent the identification of any personal information.

Specifically, the segments extracted were automatically cut into clips of exactly 500ms, extracting neighboring silence when necessary, before being uploaded on the Zooniverse platform. We will call this level *chunk*, to convey that they are part of a bigger unit (the LENA™-identified "segment"). Note that segments and chunks are collectively referred to as clips because they are parts of the longer audio recording. To allow the recovery of the original segments at later stages of the analysis, the scripts created a metadata file with the mapping between 500ms chunks and their corresponding segment. Chunks were then uploaded on Zooniverse using Panoptes, their open-source, command-line based API for data handling (Bowyer, Lintott, Hines, Allan, & Paget, 2015).

Citizen scientists could access our project, "Maturity of Baby Sounds", by logging in, or as anonymous participants. Before starting the annotation, participants were given a quick tutorial which walked them through the steps of the annotation workflow (i.e., how to play a sound, how to make a selection; see Figure 1). In this tutorial, we included one audio example for each category in order to make the classification task as smooth as possible. Further clarifications on what constitutes canonical and non-canonical sequences, as well as five audio examples for each of the categories, were available through the Field Guide, which users could access by clicking on the right side of the screen. The Field Guide could be consulted at any point of the classification without interrupting the task.

—– INSERT FIGURE 1 HERE —–

Users were asked to assign each 500ms chunk to one out of five possible categories: (1)

Canonical, (2) Non-Canonical, (3) Crying, (4) Laughing, and (5) Junk (overlapping speech, non-infant speech, silence, external sounds). Notice that the latter is narrower than the "Don't mark" category of the lab annotation routine, which was used any time there was noise or overlap that could affect acoustic analyses.

Following recommendations from the Zooniverse board, we collected five judgments per clip, rather than three as used in the laboratory. Majority agreement is thus achieved when three out of five judgments agree (rather than two out of three).

Data post-processing. An impressive total of 4,825 individual Zooniverse users provided labels for the Maturity of Baby Sounds project, of which the present data set is one part. For this project, we collected a total of 169,767 judgments provided for 33,731 500-ms chunks, corresponding to 11,980 LENA^{TM} segments. Nearly a fifth of chunks did not have at least 3 labels in agreement out of the 5 Zooniverse labels (N = 6,585, 19% of all chunks). Of the chunks without a majority agreement, 4341 (66%) contained one or two Junk judgments (out of 5), 6523 (99,9%) had at least two matching judgments (the threshold used for lab-annotated segments), and only 61 (0,01%) had 5 different judgments. Future work may explore different ways of setting the minimal requirement for convergence, but for further analyses here, we focused on the 81% of chunks that did have at least 3 labels in agreement; this represented 135,725 labels for 27,145 chunks, corresponding to 11,593 LENA^{TM} segments. As the segments average 1.12 seconds in length, this means about 3.8 hours of audio data were annotated by 8 different annotators (3 in the laboratory, 5 on Zooniverse).

Lab annotators provided judgments at the level of LENA^{TM} segments, whereas Zooniverse annotators provided judgments on 500 ms chunks, which are typically smaller than the segments. Therefore, we needed to combine chunk-level judgments to reconstruct segment-level judgments.

We did this by considering the majority label extracted from Zooniverse judgments for all chunks associated with each segment, and then following these rules: If the majority label for all chunks associated with a given segment were Junk, then the segment was labelled as Junk; then, following a hierarchy (Canonical>Non-canonical>Crying>Laughing), segments with at least one instance of each judgment were labelled as such:

- if one Canonical judgment was present, then the segment was labelled as Canonical,
- else, if a Non-Canonical judgment was present, the segment was labelled as Non-Canonical,
- else, if a Crying judgment was present, the segment was labelled as Crying,
- else, the segment was labelled as Laughing.

This is essentially the same thought process expert annotators followed when labeling segment-level audio, as each segment can represent multiple classification subtypes. LENA^{TM} segments are not equivalent to child vocalizations: A segment may contain 1 or more vocalizations, which may be of the same type or not. Figure 2 conveys the relationship between chunks and segments, and how judgments are combined into labels.

—– INSERT FIGURE 2 HERE —–

Results

Descriptive analyses. In this section, we provide descriptive analyses of our dataset. According to lab annotators, 14% of segments were Canonical, 60% Non-Canonical, 1% Laughing, and 4% Crying, with the remaining 21% being categorized as "Don't mark". Zooniverse data revealed a similar distribution: 15% Canonical, 60% Non-Canonical, 3% Laughing, 7% Crying, 15% Junk. Next, we inspected the relationship between age and child-level derived metrics, of which we had two: i) Linguistic Proportion = (Canonical+Non-Canonical)/All vocalizations (i.e., we remove Junk), and ii) Canonical Proportion = Canonical/(Canonical+Non-Canonical) (i.e., we remove Junk, Crying, and

Laughing). Figure 3 shows results of plotting our Proportions against child age. We draw the readers' attention to the fact that the children diagnosed with Angelman syndrome had overall higher levels of Linguistic Proportion than the typically-developing infants, but the opposite was true for Canonical Proportion.

—– INSERT FIGURE 3 HERE —–

Descriptive analyses on the laboratory annotations showed that correlations between the Linguistic Proportion and age differed across the groups. There was a near-zero relationship among the children diagnosed with Angelman syndrome r(8) = -0.09, CI [-0.68,0.57], p=0.80]; and a significant association among low-risk control children r(8) = 0.70, CI [0.13,0.92], p=0.02]. The Canonical Proportion exhibited non-significant developmental decreases among children diagnosed with Angelman syndrome r(8) = -0.39, CI [-0.82,0.31], p=0.26]; and marginal developmental increases among low-risk control r(8) = 0.52, CI [-0.17,0.86], p=0.13].

Using the Zooniverse annotations, we found that the association with age was very weak for children diagnosed with Angelman syndrome r(8) = -0.17, CI [-0.72,0.51], p=0.63]; whereas low-risk control children showed a significant increase with age r(8) = 0.63, CI [0.00,0.90], p=0.05]. Similarly, there were non-significant developmental decreases in the Canonical Proportion among children with Angelman syndrome r(8) = -0.41, CI [-0.82,0.30], p=0.24]; and marginal developmental increases among low-risk control children, r(8) = 0.51, CI [-0.17,0.86], p=0.13].

Main analyses

Next, we discuss the correspondence between citizen science classifications and the laboratory gold standard, at the level of individual segments. Results were visualized with a confusion matrix showing precision and recall (Figure 4): the diagonal elements show the number of correct segment-level classifications for each class while the off-diagonal elements show non-matching classifications.

—– INSERT FIGURE 4 HERE —–

These visualizations suggest that performance is moderate to good, which was confirmed via statistical analyses. The overall (weighted) accuracy is 73%, CI = [72,74], kappa is 0.53, and the Gwet's AC1 coefficient is 0.68, CI = [0.67,0.69].

Child level descriptors. Although the classification at the segment level is only moderately accurate, what we are ultimately interested in is whether citizen scientists' classifications are able to provide a reliable snapshot of children's individual development. We illustrate this association in Figure 5. Looking at all 20 children together, we found a strong positive correlation $r(18) = 0.79$, CI [0.54,0.92], p=0] between Linguistic Proportion by child from the Zooniverse and the lab annotators' data. When we split by participant group, correlations remain high: for Angelman syndrome $r(8) = 0.88$, CI [0.56,0.97], p=0.00]; low-risk control $r(8) = 0.83$, CI [0.42,0.96], p=0.00].

—– INSERT FIGURE 5 HERE —–

Similarly, a strong positive correlation is found in the Canonical Proportion $r(18) = 0.92$, CI [0.81,0.97], p=0]. When we split by participant group, correlations remain high although we do note they are somewhat smaller for the children with Angelman syndrome: $r(8) = 0.68$, CI [0.09,0.92], p=0.03]; than the low-risk control children $r(8) = 0.98$, CI [0.93,1.00], p=0.

Additional analyses

In this section, we report on explorations of under what conditions Zooniverse judgments more closely aligned with laboratory judgments. In previous work using a similar method, for instance, data from all three children from one dataset were often labeled as "Junk" (i.e., not a child's vocalization), and the data points from this corpus stood out when the authors attempted to integrate results with other corpora (Cychosz et al., 2019). A high proportion of "Junk" may indicate that automated segmentation was errorful for those children, and may be a sign that the rest of the data could be compromised as well.

We investigated this hypothesis by calculating the proportion of their data labeled as "Junk" for each individual child. There was no significant difference in the proportions of their data labeled as "Junk" for children with Angelman syndrome (M = 0.14) compared to the low-risk group (M = 0.16): Welch's t(13.35) = -0.58, p = 0.57. We therefore collapsed across groups for this exploratory analysis, and split the 20 children using a median split on the proportion of their data labeled as "Junk". Results were similar across these two post-hoc subgroups (see Figure 6). For Linguistic Proportion, the correlation across lab and Zooniverse data for the lower Junk group was r(8) = 0.86, CI [0.51,0.97], p=0.00]; and for the higher Junk group it was r(8) = 0.90, CI [0.61,0.98], p=0. For Canonical Proportion, the correlation across lab and Zooniverse data for the lower Junk group was r(8) = 0.96, CI [0.82,0.99], p=0]; and for the higher Junk group it was r(8) = 0.90, CI [0.61,0.98], p=0. Thus, it does not seem that a higher proportion of "Junk" judgments is an index of low quality data.

—– INSERT FIGURE 6 HERE —–

Above, we concluded that derived metrics integrating information across audio clips (linguistic and canonical proportion, which can be derived from segment- or chunk-level judgments) seem more promising than segment-level data (where individual segments are classified into Crying, Laughing, etc.) Notice that our derived metrics do not require matching of chunks (500 ms presented to Zooniverse participants) to segments (the original LENA^{TM} segments presented to laboratory participants), because one can calculate the proportions directly from the chunk-level judgment. As a result, there was one stage in our pre-processing that may not have been necessary, whereby we collapsed judgments across chunks associated to the same segment. We therefore repeated our analyses but instead of deriving our proportions for the Zooniverse data from the segment-level composite, we did it based on the individual chunk-level annotations. To facilitate comparison, correlations across laboratory- and Zooniverse-derived proportions are provided in Table 2.

—– INSERT TABLE 2 HERE —–

Regarding correlations with age, we found that the Linguistic Proportion was not strongly associated with age in the Angelman syndrome group: r(8) = -0.04, CI [-0.65,0.61], p=0.92], but increased with age in the low-risk control group: r(8) = 0.72, CI [0.16,0.93], p=0.02. Similarly, the Canonical Proportion did not exhibit the same pattern across the groups, with developmental decreases found among children with Angelman syndrome r(8) = -0.38, CI [-0.81,0.33], p=0.28]; and developmental increases among low-risk control r(8) = 0.57, CI [-0.09,0.88], p=0.08.

As for correlations between Zooniverse- and laboratory-derived metrics (e.g., linguistic proportion as estimated using Zooniverse coding, versus laboratory coding), we observe very similar levels of correlation: Linguistic Proportion overall r(18) = 0.80, CI [0.55,0.92], p=0] between Linguistic Proportion by child from the Zooniverse and the lab annotators' data. When we split by participant group, correlations remain high: for Angelman syndrome r(8) = 0.84, CI [0.46,0.96], p=0.00]; low-risk control r(8) = 0.79, CI [0.33,0.95], p=0.01]. As for Canonical Proportion overall r(18) = 0.95, CI [0.87,0.98], p=0; Angelman syndrome: r(8) = 0.72, CI [0.17,0.93], p=0.02]; low-risk control infants r(8) = 0.97, CI [0.88,0.99], p=0.

Since all of these results are very similar to those obtained when collapsing chunk-level judgments to generate segment-level judgments, we conclude that in the future this step may not be necessary. Instead, researchers can derive linguistic and Canonical Proportions directly from citizen scientists' chunk-level judgments (which was in fact what Cychosz et al., 2019 did).

Next, we looked at whether having more segments from each child may lead to more reliable metrics. We observed a non-significant trend [t(17.88)=348.70 and 375.50] for lower number of segments in the Angelman syndrome group (mean 348.70, range 163-599) than among low-risk control infants (mean 375.50, range 224-627), likely because of the lower volubility of the former children. Moreover, these numbers of segments are much larger than those used in previous work relying on citizen science classifications (Cychosz et al., 2019). We therefore asked whether the correlations between laboratory and Zooniverse child-level estimates are affected by how much data is extracted from each child by sampling 100 segments from all children (following Cychosz et al., 2019). We randomly sampled 100 segments from each child's data, and recalculated the Linguistic and Canonical Proportions based only on these associated judgments. We repeated this process 50 times, to assess the extent to which the association between laboratory- and Zooniverse-derived metrics varied in each random sample (e.g., the correlation between Linguistic Proportion as estimated using Zooniverse coding on the one hand, and Linguistic Proportion as estimated using laboratory coding on the other). The mean correlations across 50 runs were lower than those recovered using all segments from each child (see Table 2), particularly for the low-risk younger infants, and for Linguistic Proportion. This may indicate that, especially in some groups of infants, 100 segments may not be sufficient to capture a stable estimate of the child's Linguistic and Canonical Proportions. We also observed quite a bit of variance across the 50 runs. For Linguistic Proportion, the range of correlations found for all infants was [0.44,0.89]; for the Angelman syndrome group [0.49,0.97]; for the low-risk group [0.29,0.99]. For Canonical Proportion, the range of correlations found for all infants was [0.75,0.94]; for the Angelman syndrome group [0.79,0.99]; for the low-risk group [0.55,0.94].

Discussion

In the present study we assessed the extent to which child vocalizations from LENA™ daylong recordings can be accurately described based on classifications collected using the citizen science platform Zooniverse. Our recordings came from children diagnosed with a genetic syndrome associated with severe language disorder, as well as low-risk controls, and included a range of different ages. Excerpts of these recordings had previously been annotated by trained experts in the lab, and served as the "gold standard".

In descriptive terms, we found that category frequency was similar across laboratory and citizen science annotations. Correlations between age and our two child-level metrics revealed different developmental patterns between our sample of older children with a diagnosis of Angelman syndrome and younger low-risk controls. Given that the two groups of children differ on both of these features (age and diagnosis), we cannot empirically tease apart these two differences. Nonetheless, inspection of Figure 3 suggests different interpretations for the two derived metrics. Overall, the pattern for Linguistic Proportion is compatible with a non-linear trajectory common to the two groups, whereby the Linguistic Proportion increases rapidly up to 20 months, and hovers around 90-100% thereafter. In this case, the lack of correlation with age in the Angelman syndrome group suggests that there is little systematic variance in this metric after 20 months of age. In contrast, the two groups seem to differ in their trajectory for Canonical Proportion, with rapid increases up to 20 months in the low-risk group, and relatively stable levels of Canonical Proportion among the children in the Angelman syndrome group.

Next, we looked at agreement between laboratory and Zooniverse annotation at the level of individual segments. We found that some categories had high precision and recall – notably Non-canonical segments were detected quite accurately, and to a lesser extent the same was true for Canonical segments. This is interesting because it suggests that Zooniverse annotations may be trusted particularly for detecting these two types of speech-like segments.

In other cases, precision was fairly high but recall was lower. This was the case of the "Junk/Don't mark" category. Higher precision and low recall indicates that when Zooniverse annotations return a "Junk" judgment, this tends to be correct, but Zooniverse annotations may not detect all elements labeled as "Don't mark" in the laboratory. Inspection of confusion matrices and overall frequencies suggest that such a pattern of errors could be due to citizen scientists accepting more clips (i.e., using the "Junk" category less) than laboratory annotators, who have access to the whole vocalization. This is consistent with the fact that our lab definition of "Junk" was more stringent than the Zooniverse definition. Indeed, laboratory coders were instructed to be sensitive since the aim was also to study acoustic characteristics of the segments (which is unrelated to the current project), and therefore they would use "Don't mark" if there was *any* noise or overlapping speech that would potentially affect acoustic analyses.

Both Crying and Laughing showed moderate to good recall, but low precision. This means that a high proportion of segments detected as these two categories in the lab can be classified in the appropriate category by Zooniverse annotators, but Zooniverse annotators also put into these categories many segments that lab annotators classified in a different manner. Most saliently, there is a problematic level of confusion between Crying and Non-Canonical, whereby lab-detected Crying was almost as likely to be classified as such or Non-canonical in Zooniverse (the recall proportion was 49 and 47% respectively), and even worse, 55% of the segments Zooniverse annotators classified as Crying were actually Non-Canonical according to lab annotations. This is likely due to the fact that lab annotators had access to the whole segment, and could even listen to the context of that segment, so they could accurately interpret a segment as crying. Thus, it is possible that additional analyses, using for instance pitch or duration characteristics at the level of the whole segment rather than the chunk could improve classification performance, but for the time being, we cannot recommend Zooniverse annotations to identify crying bouts.

As for Laughing, the recall is quite good but precision is low, with confusion involving "Junk" and Non-Canonical to similar extents. We are tentative in our interpretation of this result because the category was very sparsely represented, with barely 200 segments classified as Laughing in the whole data set. We thus recommend additional work, although we expect that perhaps there will be a similar problem (and solution) as indicated for Crying, whereby determination of Laughing versus Non-Canonical will require inspection of segment-level properties.

Turning now to the child-level analyses, we were particularly interested in derived metrics because this is more typically the goal of such analyses, for instance to study potential individual or group differences, or investigate the impact of an intervention. Our results show that, despite agreement variation at the level of the segment, derived metrics were highly correlated across laboratory and Zooniverse data, meaning that variability in precision and recall did not substantially impact key outcome variables. Errors only seemed systematic for the Linguistic Proportion (and not for Canonical Proportion), and they were due to lower Linguistic Proportions from Zooniverse than laboratory data. This is due to the fact that a substantial number of Crying segments were classified as Non-Canonical by Zooniverse annotators. Nonetheless, since these errors were systematic across infants, they do not affect the study of group or individual variation.

Our exploratory analyses then dug into some methodological aspects of our data. We found that, although individuals vary widely in terms of what proportion of their data is classified as Junk by Zooniverse annotators, this is not a sign that their data as a whole is compromised, with similar levels of correlations between laboratory- and Zooniverse-derived metrics across median groups based on proportion of their data classified as Junk.

Next, we found that more data led to more reliable measurements: Specifically, our correlations based on 300-500 segments per child were markedly higher than those based on 100 segments per child, particularly when looking at our low-risk, younger participants. It is important to note that an advantage of Zooniverse, compared to laboratory annotations, is that it is easier to scale up: More data from each individual child can be annotated this way, which will help countering noise.

In sum, we believe these data show that young children's vocalizations can be accurately described by relying on the judgments of Zooniverse's citizen scientists, particularly when the goal is to describe vocalizations at a broader level than individual chunks or segments. It is important to note that from a methodological point of view, the use of crowdsourcing is not merely a question of being an ecological technique, but it can also offer benefits in the area of scientific rigor and reproducibility. As previously mentioned in the Introduction section, Zooniverse offers both an automated system of data handling as well as an enormous pool of users: access to a large-scale group of blinded participants can in itself help researchers overcome the methodological compromises that sometimes lead to biased results. Additionally, crowd-sourcing and open-science frameworks can come together to encourage replication of existing work: open-source software, shared data and citizen science platforms would make it possible, and actually easy, to run identical studies and evaluate research results independently.

Further research directions. These findings open up a new, exciting avenue for future research. Our pipeline to upload data on Zooniverse is scalable and quickly adapted to new datasets, including data on infants of different ages, learning different languages, or living in environments with different acoustic properties (e.g., rural versus urban environments). This means large existing datasets of daylong recordings, notably those hosted on Homebank (Vandam et al., 2016), can be reliably annotated in an ecological way using citizen science.

Notwithstanding our enthusiasm, we would like to point out some limitations of our results that can be addressed with more research. The first area in which more research would be welcome involves the pre-processing. We used LENA™ algorithms to detect vocalizations by the key child, but there are newer, better-performing, open-sourced alternatives that could be used instead (**???**). Still, performance is most often established on children whose development is thought to be normative, being raised in small households located in urban environments, whereas it would be important to check for accuracy in more diverse samples, where the children have a diagnosis, multiple siblings, and/or are growing up in a rural environment. Speech technology developments would also be welcome to complement citizen scientists' classifications, particularly for categories that may be hard to classify based on short samples. **remaining methodological problems with the LENA pipeline (p. XX); The team who won the challenge improved this by about 2%, primarily through gains in the laughing class obtained by adding training data (Yeh et al., 2019). Given that the laughing class is very rare, improvements in this class do not mean that performance as a whole, when all classes are weighted based on their frequency of occurrence, changes very much.**

Our results warrant additional research on two metrics we used to describe children's linguistic productions: Linguistic Proportion and Canonical Proportion. Linguistic Proportion has not been investigated extensively in previous research, and correlations with age were weaker than those for Canonical Proportion. We hope additional research is devoted to further investigating this measure as a child-level descriptor, since it may be more relevant for other disorders that are related to emotional disorders (see a review in Halpern & Coelho, 2016). Additionally, while Canonical Proportion has been studied in much previous work, we believe further research is necessary using larger scales (saliently increasing the sheer number of children studied, to have sufficient power to detect group effects) and with broader coverage (notably including children growing up in a broad range of languages and cultures, as in Cychosz et al., 2019), to make sure that assumptions regarding cross-linguistic and maturational universality are amply justified and not merely a result of having low power to detect differences, or simply sampling from similar populations.

Assuming these child-level metrics continue to be supported by such additional data, it would be interesting for future investigations to expand the analysis of Linguistic Proportion and Canonical Proportion to recover differences and similarities in language development between low-risk children and children with disorders other than Angelman syndrome. That is, it is at present unclear whether relatively lower Canonical Proportions will also be found in other populations, or whether this is more specific to Angelman syndrome. For instance, some previous work has found differences in babbling even when comparing late talkers with toddlers with no specific diagnosis (Fasolo, Majorano, & D'Odorico, 2008). We thus encourage further research inspecting Canonical Proportion in other populations, as well as methodological work that attempts to generate metrics that are similarly easy to gather at scale.

relative dearth of LENA data with atypical populations

It is also important to note that work on children with non-normative development poses in itself some challenges. Our data contained a lower number of vocalizations for children diagnosed with Angelman syndrome, even though we selected much younger low-risk control children as the comparison group. However, this limitation could be easily overcome in the future: We were limited here because we wanted to use the exact same segments that laboratory annotators had inspected. Now that this method has been validated, researchers can extract all segments that LENA™ identifies as being the key child, thus maximally using available data to get more reliable estimates, as suggested by one of our exploratory analyses.**the relatively small sample sizes of the two groups we studied (p. XX); the kinds of populations that we cannot be certain this work generalizes to (p. XX).**

Another open question we hope future work looks into is the assessment of within-participant changes. Our dataset only contained one daylong recording per child (with one exception), so other datasets in which a greater number of recordings are collected longitudinally can bring novel insights. Is the age-related decline we observed among children in the Angelman syndrome group truly a developmental decline that happens by necessity in this population, or might it reflect partially other aspects of children's experience and behavior? Additionally, longitudinal work within the same infant may be needed to track the "natural history" of vocalizations to better contextualize how these metrics should be interpreted. It is also an open question how large must effects be, at the individual level, in order to be detected with this method. For example, there is interest in using naturalistic recordings as outcome measures for clinical (behavioral or pharmacological) trials. The fact that agreement across annotators is high even in the data coming from children diagnosed with Angelman syndrome is encouraging, but it would be important to establish whether absolute levels of Canonical Proportion, and/or rates of change over longer time spans, may be good candidates for tracking the effects of any treatment.

Finally, broader generalizations of our technique could be of interest to readers of this paper. Any researcher is able to undergo the procedure to create Zooniverse projects (https://www.zooniverse.org/about/faq). For instance, a project could be set up to annotate potentially disordered speech by older adults at risk of, or diagnosed with, neurodegenerative diseases. Although it would be important to similarly validate such annotations, we believe there is strong promise for such tasks, because many neurodegenerative disorders affect local aspects of the speech, which are detected at the syllable level. In contrast, tasks that require more speech or audio context may not be well suited to citizen science platforms, because they would require playing longer clips, which may reveal identifying or sensitive information. Although citizen scientists themselves are well-intentioned, the platform could be used by others who want to exploit the system for other means.

Conclusion

In this study we validated the quality of annotations obtained through a citizen science platform, Zooniverse, as compared to a gold standard of human expert annotators. We analyzed the correspondence between annotations at the individual segment level, and the individual-child level, using Canonical Proportion and Linguistic Proportion as descriptors. We found moderate to good accuracy in the first, and strong positive correlations in the latter. We can conclude that citizen scientists are a reliable source of fast and ecological annotation of speech data, particularly when results are combined into child-level descriptors. The same methodology may be applied to several research questions in the study of language acquisition and language disorders. This finding is particularly welcome in an era when wearables open new avenues for studying human behavior and development in an ecological manner.

Acknowledgements

We are grateful to the families who contributed their data; to the Zooniverse volunteers that made this work possible; and the audience and technical committee of SLT 2020 for helpful feedback on preliminary analyses.

References

Aust, F., & Barth, M. (2017). Papaja: Create APA manuscripts with R Markdown.

Baumer, B., & Udwin, D. (2015). R markdown. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(3), 167–177.

Belardi, K., Watson, L. R., Faldowski, R. A., Hazlett, H., Crais, E., Baranek, G. T., … Oller, D. K. (2017). A retrospective video analysis of canonical babbling and volubility in infants with Fragile X Syndrome at 9–12 months of age. *Journal of Autism and Developmental Disorders*, *47*(4), 1193–1206.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Borne, K., & Zooniverse Team. (2011). The zooniverse: A framework for knowledge discovery from citizen science data. *AGUFM*, *2011*, ED23C–0650.

Bowyer, A., Lintott, C., Hines, G., Allan, C., & Paget, E. (2015). Panoptes, a project building tool for citizen science. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP'15)*. San Diego, CA.

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).

Casillas, M., & Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra*, *5*(1).

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105.

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., ... Bergelson, E. (2020). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, 1–20.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., ... Seidl, A. (2019). Canonical babble development in a large-scale crosslinguistic corpus.

Fasolo, M., Majorano, M., & D'Odorico, L. (2008). Babbling and first words in children with slow expressive development. *Clinical Linguistics & Phonetics*, *22*(2), 83–94.

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

Halpern, R., & Coelho, R. (2016). Excessive crying in infants. *Jornal de Pediatria*, *92*(3), S40–S45.

Hamrick, L., Seidl, A., & Kelleher, B. (2019). Capturing variability in early language skills among children with Angelman syndrome: A novel semi-automated approach.

Presented at the Angelman Biomarkers and Outcome Measures Alliance Scientific Meeting.

Hamrick, L., & Tonnsen, B. (2019). Measurement and characterization of early social communication in children with Angelman syndrome. In A. Sadhwani & K. Okoniewski (Eds.), *Advances in characterizing developmental trajectories in Angelman syndrome across the lifespan, Gatlinburg conference.* San Antonio, TX.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902.

McDaniel, J., Yoder, P., Estes, A., & Rogers, S. J. (2020). Predicting expressive language from early vocalizations in young children with autism spectrum disorder: Which vocal measure is best? *Journal of Speech, Language, and Hearing Research*, *63*(5), 1509–1520.

Oller, D. (2000). The emergence of the capacity for speech. Mahwah, NJ: Erlbaum.

Oller, D. K., Eilers, R. E., Neal, A. R., & Cobo-Lewis, A. B. (1998). Late onset canonical babbling: A possible early marker of abnormal development. *American Journal on Mental Retardation*, *103*(3), 249–263.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., … Warren, S. F. (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.

Overby, M., Belardi, K., & Schreiber, J. (2020). A retrospective video analysis of canonical babbling and volubility in infants later diagnosed with childhood apraxia of speech. *Clinical Linguistics & Phonetics*, *34*(7), 634–651.

Rankine, J., Li, E., Lurie, S., Rieger, H., Fourie, E., Siper, P. M., ... Kolevzon, A. (2017). Language ENvironment Analysis (LENA) in Phelan-McDermid syndrome: Validity and suggestions for use in minimally verbal children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *47*(6), 1605–1617.

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., ... others. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech* (pp. 2378–2382).

Seidl, A., Warlaumont, A., & Cristia, A. (2019). Towards detection of canonical babbling by citizen scientists: Performance as a function of clip length. In *Interspeech* (pp. 3579–3583). https://doi.org/10.21437/Interspeech.2019-1773

Semenzin, C., & Cristia, A. (2020). Zooniverse pipeline for long format recordings. Retrieved from https://github.com/LAAC-LSCP/Zooniverse

Sescleifer, A. M., Francoisse, C. A., & Lin, A. Y. (2018). Systematic review: Online crowdsourcing to assess perceptual speech outcomes. *Journal of Surgical Research*, *232*, 351–364.

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

Vandam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & Macwhinney, B. (2016). *Seminars in Speech and Language*, *37*(2), 128–142. https://doi.org/10.1055/s-0036-1580745

VanDam, M., & Yoshinaga-Itano, C. (2019). Use of the LENA Autism Screen with children who are deaf or hard of hearing. *Medicina*, *55*(8), 495.

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, *57*, 100921.

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., & Hansen, J. (2008). Signal processing for young child speech language development. In *First workshop on Child, Computer and Interaction*.

Yeh, S.-L., Chao, G.-Y., Su, B.-H., Huang, Y.-L., Lin, M.-H., Tsai, Y.-C., … others. (2019). Using Attention Networks and Adversarial Augmentation for Styrian Dialect, Continuous Sleepiness and Baby Sound Recognition. In *INTERSPEECH* (pp. 2398–2402).

Table 1

Comparison of recall percentages obtained with the baseline algorithm created by the ComParE team, in the context of the ComParE 2019 BabySounds subchallenge (Schuller et al., 2019), LENA™ labels (note LENA™ does not distinguish between Canonical and Non-canonical), and those obtained in this study through Zooniverse annotations. Label frequency indicates the prevalence of the relevant label (Crying, Laughing, etc.) in each dataset (the same Lab dataset was used for both LENA™ and Zooniverse). UAR stands for unweighted average recall, WAR for weighted average recall (which takes into account label frequency).

<table>
<tr><th></th><th>C2019B</th><th>LENA</th><th>Zoon.</th><th>C2019B</th><th>Lab</th></tr>
<tr><td>Crying</td><td>70.6</td><td>78.0</td><td>49.0</td><td>669</td><td>581</td></tr>
<tr><td>Laughing</td><td>41.5</td><td>12.0</td><td>56.0</td><td>149</td><td>137</td></tr>
<tr><td>Non-can.</td><td>24.1</td><td rowspan="2">73.0</td><td>77.0</td><td>4485</td><td>5512</td></tr>
<tr><td>Canonical</td><td>66.4</td><td>59.0</td><td>1426</td><td>1715</td></tr>
<tr><td>Junk</td><td>67.3</td><td>1.0</td><td>44.0</td><td>4575</td><td>2169</td></tr>
<tr><td>UAR</td><td>54.0</td><td>41.0</td><td>57.0</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>WAR</td><td>49.9</td><td>57.0</td><td>65.0</td></tr>
</table>

Table 2

Pearson correlation coefficients across metrics derived from laboratory and Zooniverse annotations in the Angelman syndrome (AS) group data, low-risk (LR) group data, or for all children together (all) in three ways. First, All seg indicates that Zooniverse annotations at the chunk level were first combined at the segment level. Second, Chunks indicates that they were analyzed directly at the chunk level. Both of these are based on all the data. Third, rows labeled 100 seg indicate that laboratory- and Zooniverse-derived metrics were based on only 100 segments (median over 50 runs in which 100 segments were randomly selected from each child).

Linguistic Proportion	Canonical Proportion
0.878	0.681
0.828	0.984
0.794	0.920

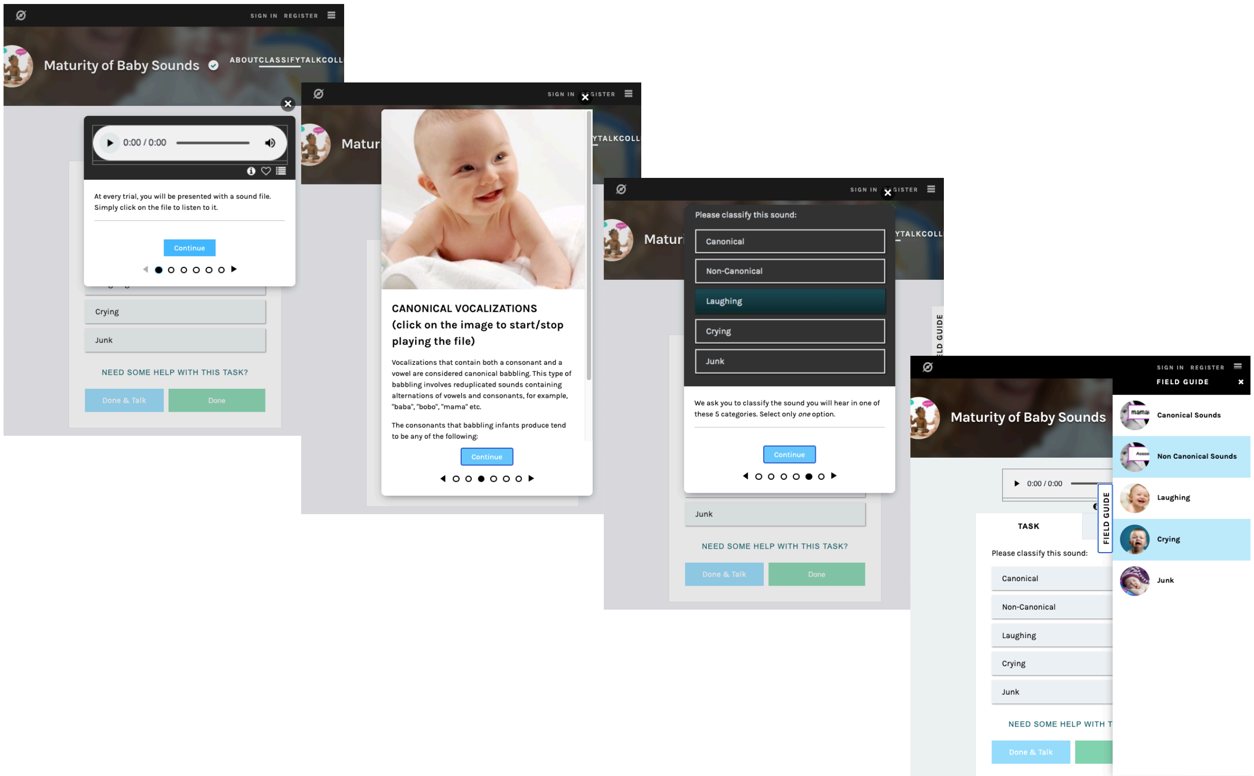

Figure 1. Screen captures of the 6 stages of the tutorial, three of which are shown here: The initial explanation of the trial, examples of canonical vocalizations, and the classification choice the user needed to make; as well as the Field Guide, which remained accessible on the right of the screen throughout the session, and which allowed users to revise examples of the different categories.

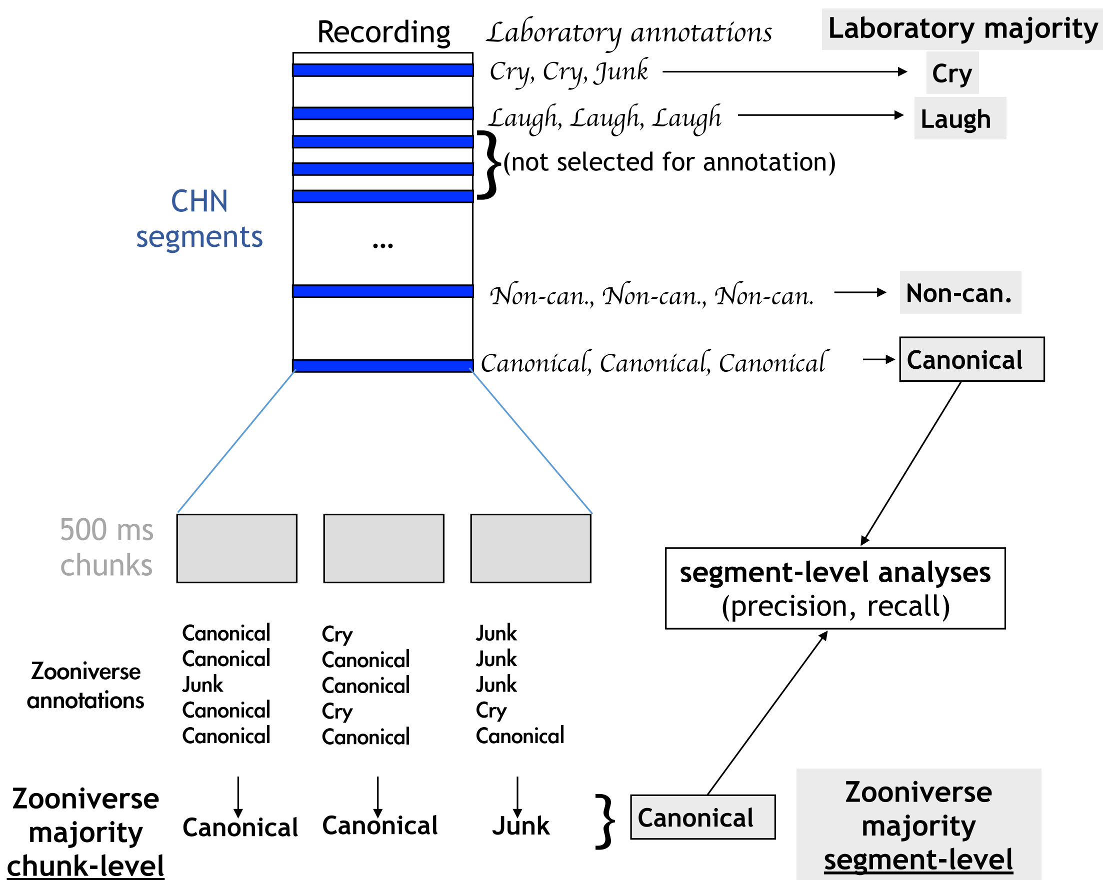

Figure 2. Correspondence between LENA-identified segments (which can contain one or more vocalizations) and chunks. Segments and chunks are collectively referred to as clips because they are parts of the longer audio recording. Notice that clips are first annotated by multiple people, and these judgments are then combined into a single label based on simple majority. To compare laboratory and Zooniverse annotations at the level of segments, chunk-level judgments are combined using the hierarchy Canonical>Non-canonical>Crying>Laughing.

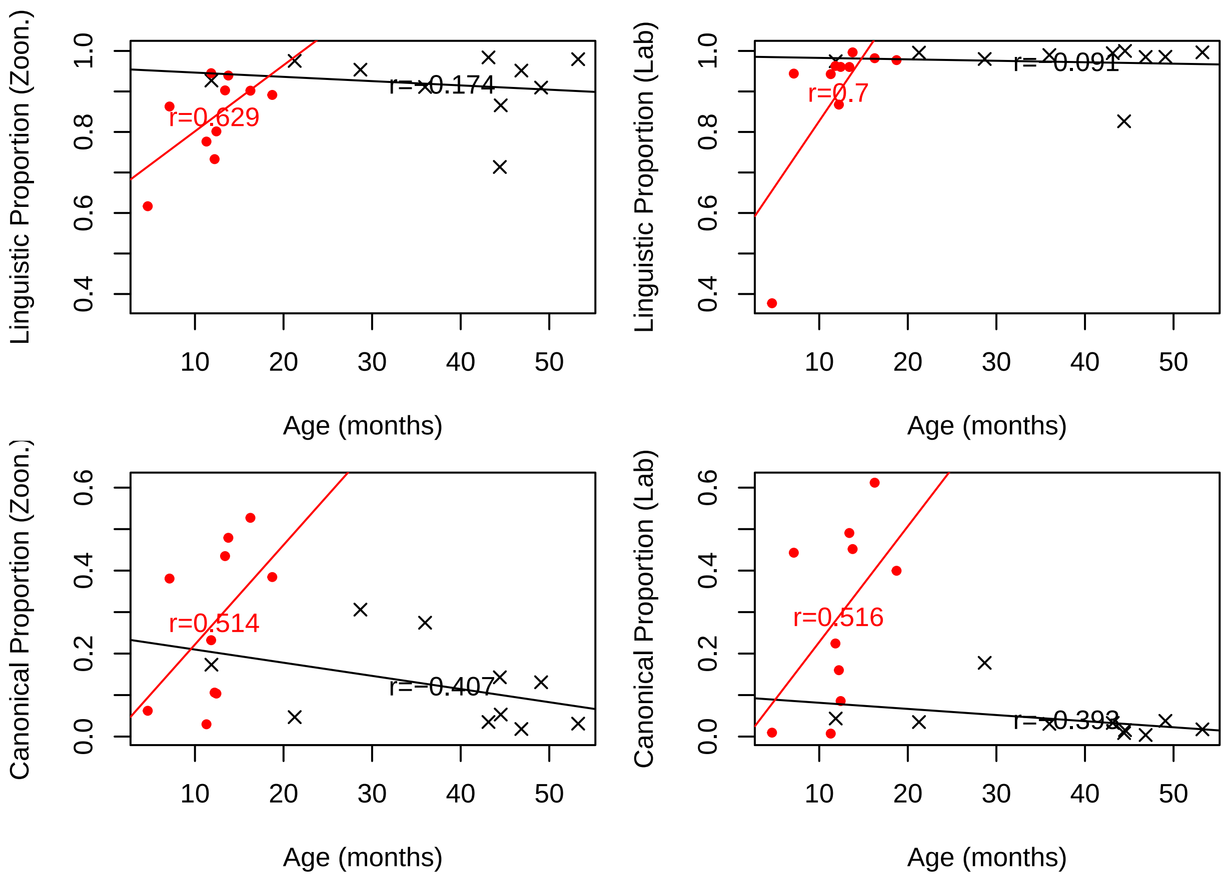

Figure 3. Correlations between child-level descriptors and age as a function of metric (Linguistic Proportion in the top row, Canonical Proportion in the bottom row), annotation method, and child group (red=low-risk, and black= Angelman syndrome).

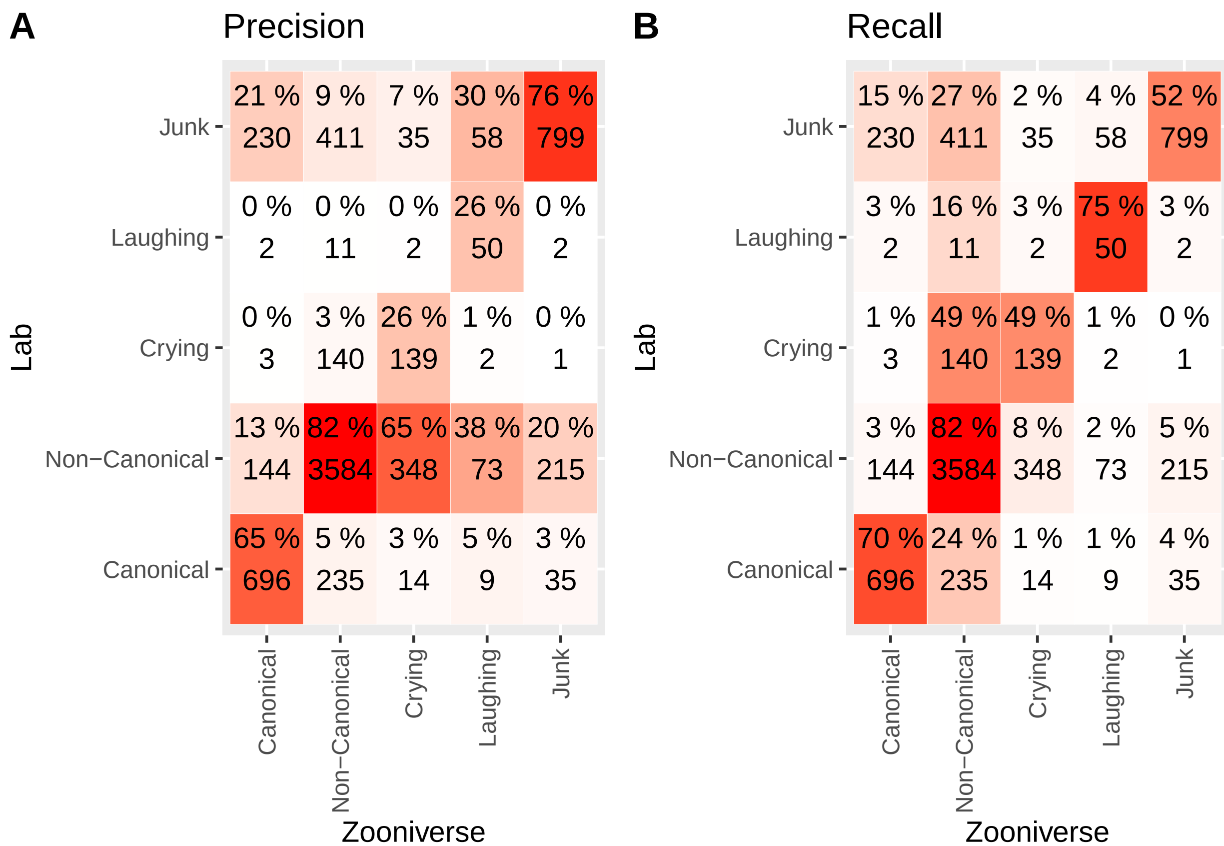

Figure 4. Number of segments as a function of their majority label from laboratory annotations (rows) versus Zooniverse annotations (columns). Percentages (and shading) indicate precision (left) and recall (right). For precision, the percentage indicates what percentage of the segments that have a given majority label according to Zooniverse data are attributed a given label in Lab data (i.e., columns add up to 100%). For recall, the percentage indicates what percentage of the segments that have a given majority label according to Lab data are attributed a given label in Zooniverse data (i.e., rows add up to 100%).

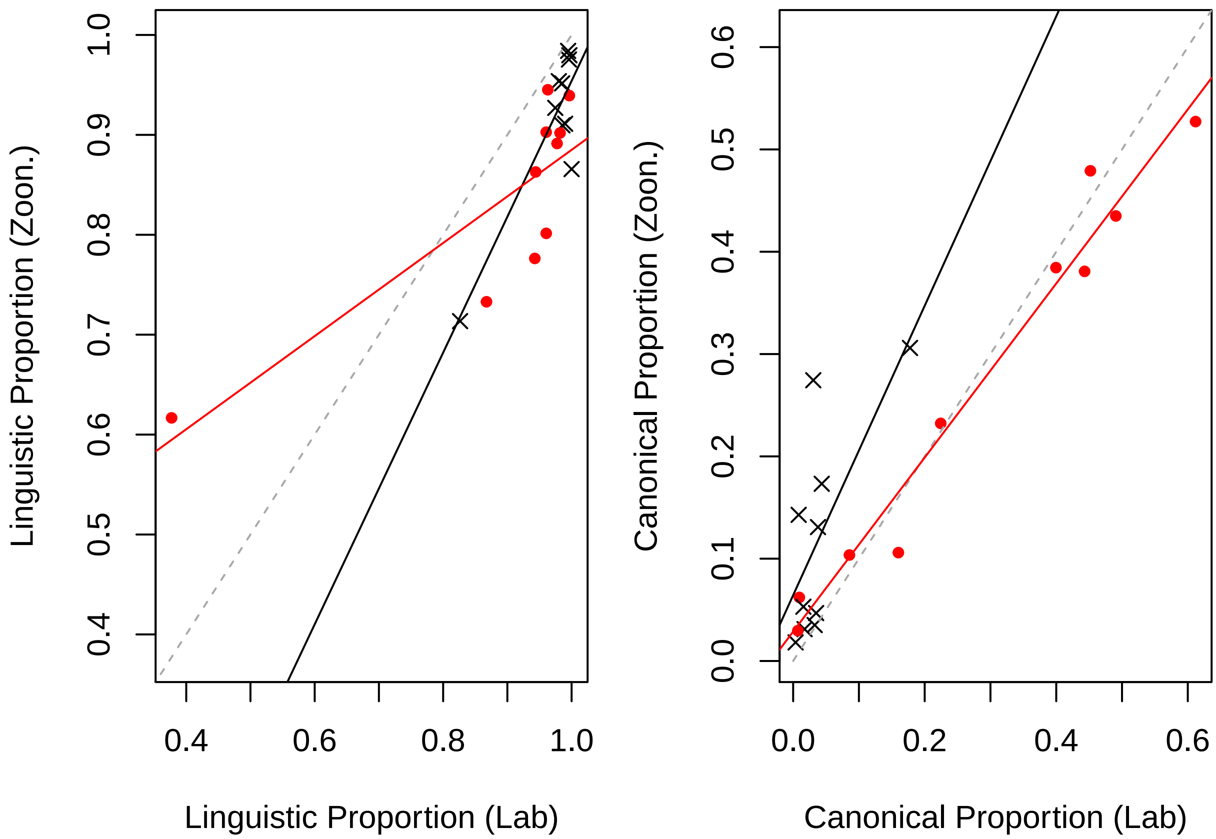

Figure 5. Correlations between child-level descriptors derived from Zooniverse data (y axis) versus lab data (x axis) as a function of child group (red=low-risk, and black=Angelman syndrome).

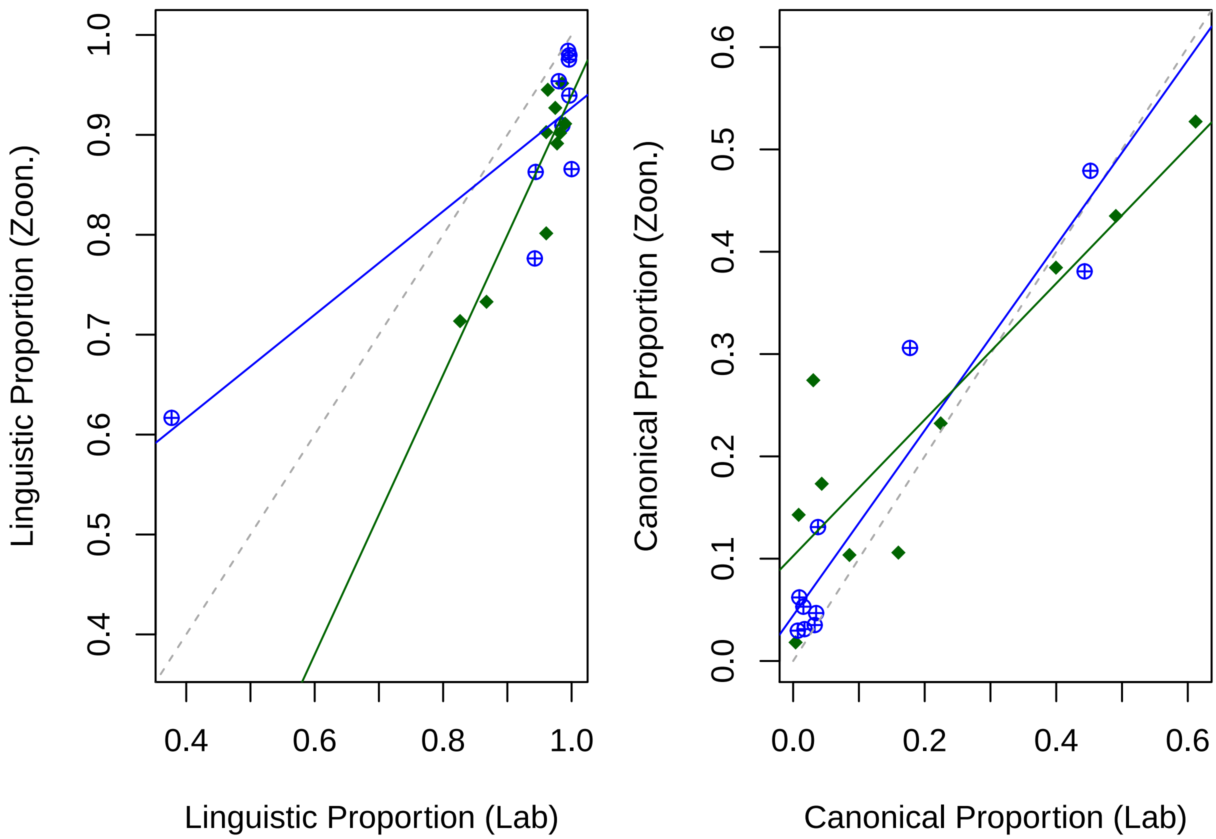

Figure 6. Correlations between child-level descriptors derived from Zooniverse data (y axis) versus lab data (x axis) as a function of whether children's Junk proportion was higher or lower than the median for all children (blue=lower, and green=higher).