Everyday language input and production in 1001 children from 6 continents

Elika Bergelson[a,1], Melanie Soderstrom[b], Iris-Corinna Schwarz[c,n], Caroline F. Rowland[d,e,f], Nairán Ramírez-Esparza[g], Lisa Rague Hamrick[h], Ellen Marklund[c], Marina Kalashnikova[i,o], Ava Guez[j], Marisa Casillas[d,f,k], Lucia Benetti[l], Petra van Alphen[m], and Alejandrina Cristia[j,1]

[a]Harvard University, Department of Psychology; [b]University of Manitoba, Department of Psychology; [c]Stockholm University, Department of Linguistics; [d]Max Planck Institute for Psycholinguistics, Language Development Department; [e]Radboud University, Donders Centre for Brain, Cognition and Behaviour; [f]ARC Centre of Excellence for the Dynamics of Language (CoEDL); [g]University of Connecticut, Psychological Sciences; [h]Purdue University, Department of Psychological Sciences; [i]Basque Center on Cognition Brain and Language (BCBL); [j]PSL University, Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS; [k]University of Chicago, Comparative Human Development Department; [l]Ohio State University, School of Music; [m]Royal Dutch Kentalis; [n]Stockholm University, Department of Special Education; [o]Ikerbasque, Basque Foundation of Science

This manuscript was compiled on August 23, 2023

Language is a universal human ability, acquired readily by young children, who otherwise struggle with many basics of survival. And yet, language ability is variable across individuals. Naturalistic and experimental observations suggest that children's linguistic skills vary with factors like socioeconomic status and children's gender. But which factors really influence children's day-to-day language use? Here we leverage speech technology in a big-data approach to report on a unique cross-cultural and diverse data set: >2,500 day-long, child-centered audio-recordings of 1,001 2- to 48-month-olds from 12 countries spanning 6 continents across urban, farmer-forager, and subsistence-farming contexts. As expected, age and language-relevant clinical risks and diagnoses predicted how much speech (and speech-like vocalization) children produced. Critically, so too did adult talk in children's environments: Children who heard more talk from adults produced more speech. In contrast to previous conclusions based on more limited sampling methods and a different set of language proxies, socioeconomic status (operationalized as maternal education) was not significantly associated with children's productions over the first four years of life, and neither were gender or multilingualism. These findings from large-scale naturalistic data advance our understanding of which factors are robust predictors of variability in the speech behaviors of young learners in a wide range of everyday contexts.

infancy | human diversity | language | socioeconomic status | speech

Typically-developing children readily progress from coos to complex sentences within just a few years, leading some to hypothesize that the universal language abilities of humans develop uniformly, with only incidental effects of individual- or group-level variation (1). And yet, studies using a variety of proxies for language development find some evidence of such variation in early language skills, with differences reported between girls and boys (2), as well as those raised in socioeconomically privileged compared to disadvantaged households (3, 4).

However interesting, these studies tend to rely on Western-centric samples and methods, and may not reflect everyday language use in children. Moreover, prior work often stops after only considering individual predictors in a binary way (i.e. do they significantly impact language development or not), while failing to ask the more informative question of *how large their relative impact is* (5), especially in freely-occurring, everyday speech behavior.

Recent research on mice and whales shows the promise of machine learning for examining everyday animal behavior (6, 7). We leverage advances in wearables and machine-learning-based speech technology to catalyze a similar breakthrough in language development research. Our dataset is comprised of >40,000 hours of audio from >2,500 days in the lives of 1,001 2- to 48-month-olds from 6 continents and diverse cultural contexts (Figure 1). Within this dataset, we focused on the *amount* of speech or speech-like vocalization young children produce in their everyday life. Critically, these automatically-extractable "quantity" measures correlate robustly with gold-standard "quality" measures of children's language skills and knowledge, like vocabulary estimates (see SI1D for relevant evidence) (4).

We query and compare the effects of two types of factors. First, there are factors with undeniable effects on early lan-

Significance Statement

Harnessing a global sample of >40,000 hours of child-centered audio capturing young children's home environment, we measured contributors to how much speech 0–4 year olds naturally produce. Amount of adult talk, age, and normative development were the sole significant predictors; child gender, socioeconomic status, and multilingualism did not explain how often children vocalized, or how much adult talk they heard. These findings (strengthened by our validation of existing automated speech algorithms) open up new conversations regarding early language development to the broader public, including parents, clinicians, educators, and policymakers. The factors explaining variance also inform our understanding of humans' unique capacity for learning, and potentially large-scale applications of machine technology to everyday human behavior.

EB, MC, and AC developed the initial conceptualization of the project and recruited corpus owners and co-authors. EB, MC, and AC curated the meta-corpus and meta-data and prepared them for analysis. EB and AC prepared materials for and/or led group decision-making. EB, MS, CR, NRE, AG, MC, LB, PvA, and AC contributed to the decision-making on the analytic approach, including selection of exploratory and confirmatory sets, selection of variables, identification of hypotheses and/or specification of models. AC, EB, and AG drafted the preregistrations. AC, EB, and AG designed and implemented the analyses. EB, CR, NRE, LRH, MK, and LB conducted and synthesized literature reviews on key topics related to the decision-making regarding literature review, hypotheses, and analyses. EB, EM, ICS, CR, LRH, MS, NRE, MK, MC, PvA, and AC provided corpus data and meta-data. See acknowledgments for non-author data contributors. EB, AC, and MS contributed to the initial manuscript draft writing. EB, MK, MC, and AC contributed to visualizations. EB, MC, and AC revised and responded to feedback and informal peer-review. MS, CR, LRH, LB, EB, AC, EM, ICS, and PvA contributed to supplementary materials, Open Science Framework project page and/or other documentation. Note: Other than first and last authors, middle authors are listed in reverse alphabetical order.

The authors have no conflict of interest to declare.

[1] To whom correspondence should be addressed. E-mail: elika_bergelson@fas.harvard.edu, alecristia@gmail.com

guage production, namely, child age and language-relevant clinical risks and diagnoses. Second, there are individual- and family-level factors that are reported to correlate with variability in early language skills: socioeconomic status (SES; operationalized here as maternal education; SI2B), gender, language input quantity, and multilingual background. Because small and homogeneous samples make universal claims more questionable, a key novel contribution of this work is its benchmarking of the level of stability and variability of everyday language use in a heterogeneous, richly diverse participant sample.*

Measuring Diverse, Real-life Language Use. Language skills and knowledge are not directly observable. As a result, all studies use a proxy when estimating them in individual children. These proxies have variable validity and predictive power relative to other measures, both concurrently and predictively, and likely vary in the extent to which they reflect children's everyday language behavior. For instance, parental report measures are indirect and—especially for receptive knowledge—can be difficult for caretakers to estimate (9), even in relatively homogeneous Western-centric contexts.

Here, we adopt a very different approach. We employed the LENA™ system, which captures what children hear and say across an entire day through small wearable recorders (10); this ecologically-valid sampling method reduces observer effects relative to, e.g., shorter video recordings (11). The LENA™ system uses standardized algorithms that estimate who is speaking when, alongside automated counts of adult and child linguistic vocalizations (4) (see definition and validation in SI1C:E). The resulting LENA™ measures correlate with and predict other measures of language skills in children with and without clinical risks or diagnoses, as revealed by manual transcriptions, clinical instruments, and parent questionnaires (12, 13). We use LENA™'s validated, automated estimates to derive our measures of everyday language use: adult talk and child speech (see detailed motivation in SI3B). We define **child speech** as the quantity of children's speech-related vocalizations (e.g., protophones (14), babbles, syllables, words, or sentences, but not laughing or crying) per hour, and **adult talk** as the number of near and clear vocalizations per hour attributed to adults (both as detected by LENA™'s algorithm; see Methods). Assuaging concerns that these measures are merely capturing chattiness or repetition, both have a $\geq .7$ correlation with measures of lexical diversity and language "quality": our child speech measure correlates with vocabulary in an independent sample, and the adult talk measure correlates with the number of word types from manual transcription in a subset of the data (SI1D).

Capitalizing on this standardized and deidentified numeric output, we solicited LENA™ datasets that researchers had previously collected to study mono- and multilingual children (i.e. those learning >1 language) in urban, farmer-forager, and subsistence-farming contexts worldwide (Figure 1). This resulted in a dataset reflecting the state of current knowledge in ecologically-valid speech samples from children's daily lives (SI3A; see Methods for more sample details).

The dataset includes children from wide-ranging SES backgrounds, based on maternal education levels spanning from no formal education to advanced degrees (SI2B). This SES proxy was selected not only because it was available in all 18 corpora (only 3 had alternative SES proxies), but most importantly because it is the most commonly employed SES proxy in language acquisition research, as established in meta-analyses (15, 16). This allows our findings to inform ongoing discussions. Theories of how SES relates to children's language development have proposed a wide range of pathways in which maternal education is predictive of children's language experiences, including the connection between maternal education and the tendency to employ verbal over physical responsiveness (17), the diversity in mothers' vocabulary (18), and the frequency of verbally-rich activities (19). Maternal education also correlates highly with other SES proxies (e.g. r=.86 in a study of children growing up in 10 European or North American countries, 20), suggesting it may also indirectly pick up on other pathways linking SES to language development, through e.g. differential access to resources and nutrition, or exposure to stress perinatally (21). At the same time, we recognize that comparing a variable like education across countries, although commonly done (22), is not straightforward. Therefore, we supplement our pre-registered approach with numerous exploratory checks and analyses examining alternative implementations (SI3G:H described further below).

Crucially, by including children aged 2 to 48 months, we span a wide range of linguistic skills, allowing us to better capture the effects of our variables over a broad span of development within our socio-culturally and geographically broad-ranging participants. We also include children with a variety of diagnoses of language delays and disorders, as well as those at high risks for them (see Methods & SI2A for definitions and detailed justification). Such children's language development is by definition *non-normative*. Thus, age and non-normative status provide useful yardsticks for considering the significance and effect size of other child- and family-level factors (SES through maternal education, child gender, mono- vs. multilingual status, and how much adults talk to and around the child). That is, if a factor (e.g., gender) has an effect far smaller than that of age or non-normative development, it would suggest that individual differences within it are relatively limited in their connection to everyday language use. If the effects are comparable in size, it would instead suggest that the amount of speech humans produce in everyday interactions is undergirded by substantial and structured individual differences, rather than striking uniformity. Given that effects could vary as a function of child age, we make sure to include key interaction terms. For instance, we can expect age to interact with adult talk if (as anticipated) older children are more sensitive to adults' talking to them than younger ones.

Predicting Children's Speech Production. We employed a hypothesis-testing approach: In a two-step preregistration, we first established exploration and confirmation data subsets (see Methods and SI3A for detailed explanation, and SI3D:E for the procedure used to derive pre-registered hypotheses and analyses). We then leveraged the held-out confirmation subset to answer our key question: **How well do specific individual- and family-level factors predict variation in how much speech young children produce?** At stake in these analyses is *whether* systematic differences in children's

*While these data collectively span living circumstances, geography, and family structure, some data donors were concerned that highlighting differences when minoritized communities are involved poses ethical challenges, in terms of honorable representation and potential harm. Individual data stewards are actively engaging in richer descriptions of included samples (see SI5), which may enable future work on meaningful population-level differences (e.g., 8).

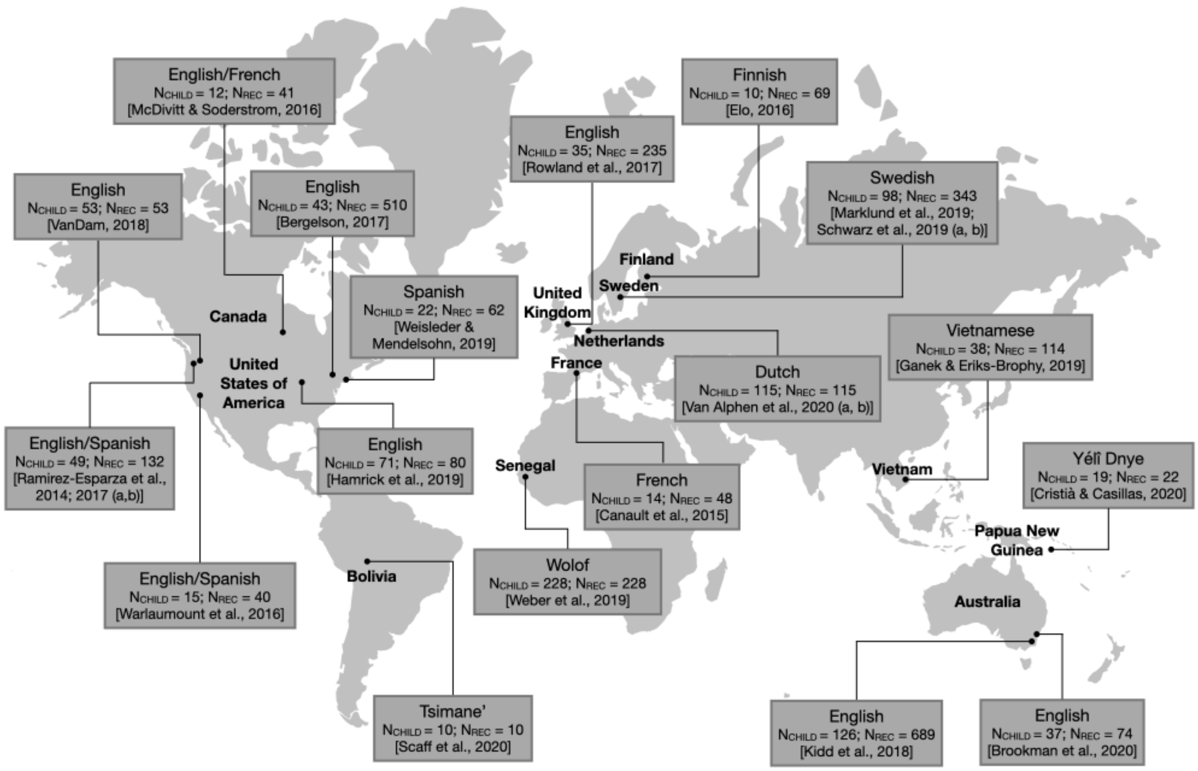

Fig. 1. Geographical location, primary language, number of children (Nchild), number of recordings (Nrec) and data citation for each corpus.

Table 1. Model results predicting child speech. q-values show FDR-corrected p-values.

	β	SE	q	
Intercept	0.109	0.128	.681	
Child Gender(Male)	0.026	0.051	.852	
SES(<H.S.(1))	0.001	0.111	.991	
SES(H.S.(2))	-0.033	0.115	.932	
SES(B.A.(4))	-0.064	0.079	.681	
SES(>B.A.(5)	-0.002	0.090	.991	
Control	-0.085	0.029	.035	*
Norm	-0.220	0.087	.036	*
Adult Talk	0.260	0.037	<.001	*
Age	0.647	0.024	<.001	*
Mono	0.045	0.095	.852	
Norm × Adult Talk	-0.005	0.063	.991	
Norm × Age	-0.217	0.051	<.001	*
Adult Talk × Age	0.125	0.022	<.001	*
Adult Talk × Mono	0.092	0.072	.45	
Mono × Age	-0.048	0.056	.681	
Norm × Adult Talk × Age	0.019	0.043	.852	
Mono × Adult Talk × Age	0.137	0.065	.094	

Note. Betas show deviation from the following baseline levels: Child Gender: female; SES: some university(3); Norm: Norm(ative development); Mono: Mono(lingual). SES = child SES based on maternal education (<H.S.(1) = less than high school, H.S.(2) = high school, B.A.(4) = college degree, >B.A.(5) = advanced degree); Control = overlap rate control; Adult Talk = adult vocalization count rate.

lives have measurable links to their language production, and if so, what the *strength* of these relationships is both overall, and in relation to one another (see Table 1 for results[†]).

As expected, we found that older children produced more speech than younger ones (ß=0.647, SE=0.024). Children with non-normative development produced less speech than children with normative development (ß=-0.22, SE=0.087)[‡], an effect that strengthened with age (ß=-0.217, SE=0.051; see Figure 2B). This is expected because for some groups in our non-normative subset (e.g. those with familial risk of a speech impairment) older children are more likely to have an actual diagnosis (as opposed to risk factor) than younger ones (see SI2A for details on non-normative classification).

Our results further revealed that young children's speech production correlated with the amount of adult talk they heard (ß=0.26, SE=0.037). This correlation strengthened with age (ß=0.125, SE=0.022; see Figure 2A), perhaps because variation in adult talk rate has less effect on infants (whose early babbles occur frequently even when infants are alone, 14). The effect of adult talk is a substantial one. Taking the effects of age and normativity as convenient (but unrelated) gauges for what counts as a considerable effect, we see that the effect size of adult talk is about a third of that for age and similar to that for normativity (adult talk: 0.26; interaction adult talk by age: 0.125; age: 0.647; non-normative development: -0.22; interaction non-normative by age: -0.217; all effect size betas expressed as SDs).

To provide these results in more intuitive units, we fit the same model centering variables without scaling. Children

[†] All ßs in Tables and text are based on treatment-coded models. See SI3H for sum-coded models, which give the same pattern of results.

[‡] The normativity estimate is negative because normative development is the baseline.

produced 66 more vocalizations per hour with each year of life. For every 100 adult vocalizations per hour, children produced 27 more vocalizations; this effect grew by 16 vocalizations per year. Relative to infants with typical development, those with non-normative development produced 20 fewer vocalizations per hour; this difference grew by 8 vocalizations per year.

Surprisingly, and in contrast to previous results using smaller and less diverse datasets and/or other language proxies, we found that child gender, SES (indexed here by maternal education), and monolingual status did not explain significant variation in child speech. As our raw data figures and model outcome results show, these null effects hold both when considering covariates (as in our model; Table 1) and when considering these variables individually (as in Figure 3; SI3F, 3G, 3H). In our full model controlling for other variables (Table 1), the largest estimate for main effects or interactions involving child gender, SES, and monolingual status was about half of that for normativity, and one-sixth of that for age; none reached thresholds for statistical significance.

While our models are well-powered to estimate associations of child speech with age, normativity, adult talk, gender, SES (as measured by maternal education), and monolingual status, this is predicated upon pooling the data and accounting statistically for corpus- and child-level variance via random effects, as described in Methods. This makes it beyond this paper's scope to analyze language or population/cultural differences in detail, i.e. in a way that might allow the consideration of additional, culture-specific variables (hence their omission in Figs 2–3); see SI5 for citations to research on individual datasets, some of which tackle such differences directly.

Noting that the results above have the strongest inferential value thanks to being pre-registered, we also addressed certain alternative hypotheses and interpretations that could have rendered our conclusions unjustified through a series of follow-up analyses. These checked for robustness of our key results with different operationalizations and statistical implementations of SES, when considering only children under or over 18 months, when considering causal paths, and when incorporating speech from other children as a predictor; our key results held in all cases (SI3H).

We highlight here the results that may run most counter to many readers' assumptions, namely, that in this large sample, SES (indexed by maternal education) does not come out as a significant predictor of child speech. This conclusion held when declaring SES as an ordinal and as a continuous variable based on levels or years of maternal education, when binarizing SES levels based on individual countries' average education completion rate, and when declaring a random slope for SES within corpus (which allows SES effects to vary across corpora). Some readers may wonder whether there were some corpora for which SES did matter. If so, the analysis with random SES slopes by corpus would have indicated this, but it did not (SI3H). The relationship between SES and child speech was weak and inconsistent across corpora (as evident in Fig. 4).

Perhaps most convincingly, results also held when constraining our analysis to our largest homogeneous subset, the North American subsample (642 daylong recordings from 206 infants in 7 corpora; SI3G). We essentially replicated the full-sample results in this subsample: adult talk and age were significant predictors, whereas gender and SES (based on maternal education) were not. The significant adult talk × age interaction also replicated. The main effect of normativity did not, likely because normativity's interaction with age was larger than in the full-sample analysis. Finally, we also tested whether removing the adult talk variable would result in an SES effect, i.e. testing whether adult talk was absorbing variance that would otherwise be accounted for by SES. This was not the case: Removing the adult talk predictor, SES still does not account for significant variance in child speech in our analysis. A central contribution of this work is thus the clear lack of evidence we find for effects of SES (under several operationalizations focused on maternal education), on how much speech young children produce in day-to-day life.

Another potential concern is that our conclusions hinge on the use of LENA™'s particular algorithm; they do not. The findings above successfully replicate in the subset of data for which data stewards were able to share raw audio (11/18 corpora), which was analyzed with a wholly different algorithmic approach, the Voice Type Classifier or VTC (Methods; SI3F).[§] Yet another worry is that our focus on adult talk may mask other important contributions to children's language experiences, for instance, speech from other children. Testing this in a supplemental analysis, we confirm that the level of association found between adult talk and children's speech was unaffected by including other children's talk measured by LENA as a predictor variable (SI3H), confirming that our key conclusions hold when factoring this other source of input in.

Finally, we also ran a model predicting adult talk (rather than child speech). The amount of adult talk was not significantly predicted by SES, child age, gender, monolingual or normative status (Table 2, Figure 3E:H; SI3G:H). Importantly, these null results replicated in the North American subset (SI3G) as well as in every other alternative analysis we attempted (SI3H). Together, these analyses suggest that the relationship we find between adult talk and child speech in the child speech models is not attributable to child- or family-level factors affecting adult talk.

Speech and Other Early Vocal Behavior. While our central query concerned variability within early speech production, we conducted a further descriptive analysis examining how much of children's vocalizations were speech or speech-like, as opposed to the two other classes of LENA™-identified vocalizations: crying and vegetative sounds (e.g. burps, hiccups). We examined these vocalization types as a function of age, monolingual status, and normative status. As Figure 2C shows, for children with normative development, the proportion of vocalizations that were speech increased from just over half to the vast majority over 2–48 months. In contrast, the crying proportion fell steeply over the same period, from nearly half of vocalizations to a small fraction of them; the proportion of vegetative sounds was low and constant. Convergent with our speech analyses, monolingual status did not alter these patterns but normative status did: While the same overall patterns held for children with non-normative development, their decrease in crying and increase in speech production with age was less steep (see Figure 2C).

As with more narrowly-defined non-normative populations (e.g. children with Autism Spectrum Disorder (23)), we find clear divergences in language trajectories in our normative vs. non-normative samples. This is notable because (a) our

[§]VTC too has been robustly validated relative to various gold standard manual measures (SI1E)

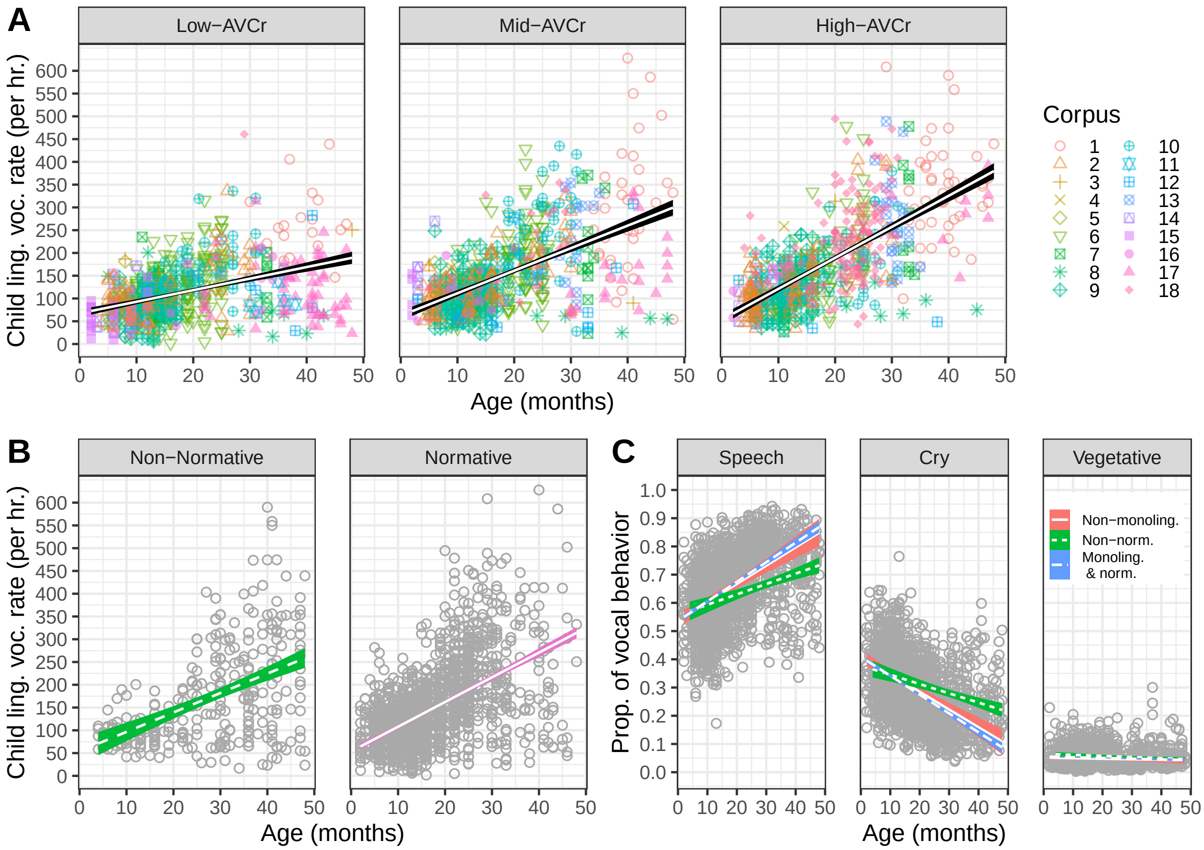

Fig. 2. Effects of adult talk, child age, and normative development on children's speech production. Points show each daylong recording; lines show linear regression with 95% Confidence Intervals (CI). Child speech is quantified as child linguistic vocalization rate; adult talk as adult vocalization count rate (AVCr). **A**: Child speech by age, split by low/mid/high tertiles of adult talk. Lines depict significant adult talk × age interaction. Color-shape combinations show each unique corpus, numbered to preserve anonymity. **B**: Child speech by age and normative status. Lines depict significant age × normative status interaction. **C**: Proportion of vocal behavior classified as speech, cry, or vegetative, by age. Line type/color indicate monolingual and normative statuses. N.B. Monolingual normative CI (blue) falls fully within that for multilingual children (pink) for all 3 types of vocal behavior, highlighting these groups' equivalent patterns.

Table 2. Model results predicting adult talk (i.e. adult vocalization count rate). q-values show FDR-corrected p-values.

	β	SE	q
Intercept	-0.100	0.160	.778
Child Gender(Male)	0.174	0.148	.547
SES(<H.S.(1))	0.239	0.173	.547
SES(H.S.(2))	-0.015	0.194	.939
SES(B.A.(4))	0.148	0.131	.547
SES(>B.A.(5)	0.098	0.150	.778
Control	0.084	0.055	.547
Norm	0.013	0.103	.939
Age	-0.030	0.029	.547
Mono	-0.028	0.112	.939
Gender(Male) × SES(<H.S.(1))	-0.375	0.196	.547
Gender(Male) × SES(H.S.(2))	-0.263	0.252	.547
Gender(Male) × SES(B.A.(4))	-0.220	0.176	.547
Gender(Male) × SES(>B.A.(5))	0.016	0.201	.939
Norm × Age	-0.076	0.060	.547
Mono × Age	0.035	0.068	.804

Note. None of the variables in our model predicted adult talk. All abbreviations and baselines as in Table 1.

non-normative sample is heterogeneous (SI2A) and (b) as 2–48-month-olds, many children with non-normative classifications here were at risk of (but not yet diagnosed with) language delays or disorders. Automated speech analyses in naturalistic recordings thus hold promise for future research into early diagnostics (24, 25).

Adult Talk and Child Speech. Children who heard more adult talk produced dramatically higher rates of speech, and this effect increased with age. This result feeds into ongoing theoretical debates regarding the relevance of individual differences (26). Although we cannot infer causality from our correlational data, it is useful to consider possible causal paths that could in principle have led to our results. A correlation between child speech and adult talk is compatible with at least three explanations: (1) Children who produce more speech *elicit* more talk from adults; (2) Language-dense environments *lead* children to produce more speech; or (3) A third variable causes increases in both adult talk and child speech.¶

Our model predicting adult talk (see Table 2) can be brought to bear on Explanation 1. If children talking more elicited more talk from adults, then we would have expected to

¶Our analyses suggest that one such potential third variable, differences in activities across recordings, is not a likely candidate for the correlation between child speech and adult talk (SI4).

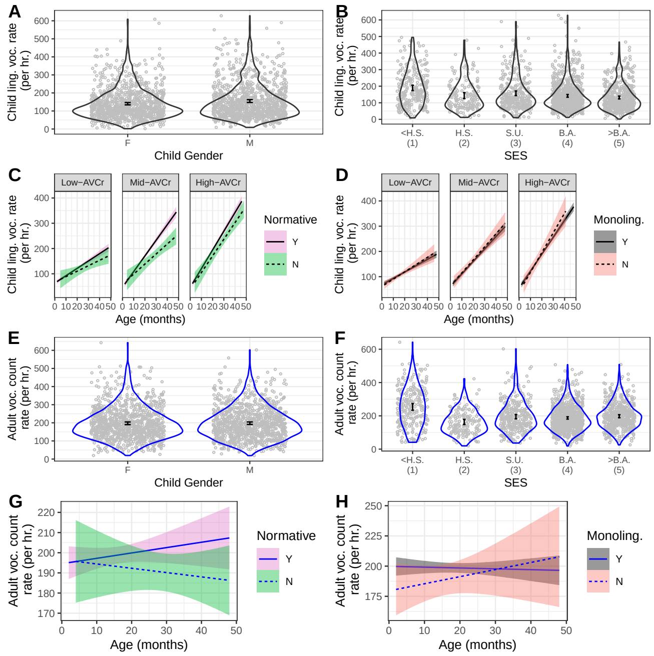

Fig. 3. Factors that do not predict child speech or adult talk. Points = individual recordings, jittered horizontally. Lines = linear fit with 95% Confidence Intervals. Error bars = 99% bootstrapped CIs of sample means. Child speech is quantified as child linguistic vocalization rate; adult talk as adult vocalization count rate (AVCr). **A & B**: null effects of child gender (**A**) and socioeconomic status (SES) (**B**) on child speech. **C**: null 3-way effect of normative development × adult talk × age (N.B.: normative × age and adult talk × age are significant; see Fig. 2). **D**: null 3-way effect of age × adult talk × monolingual status. **E** and **F**: null effects of child gender (**E**) and SES (**F**) on adult talk. **G & H**: null effect of normative development (**G**) and monolingual status (**H**) on adult talk.

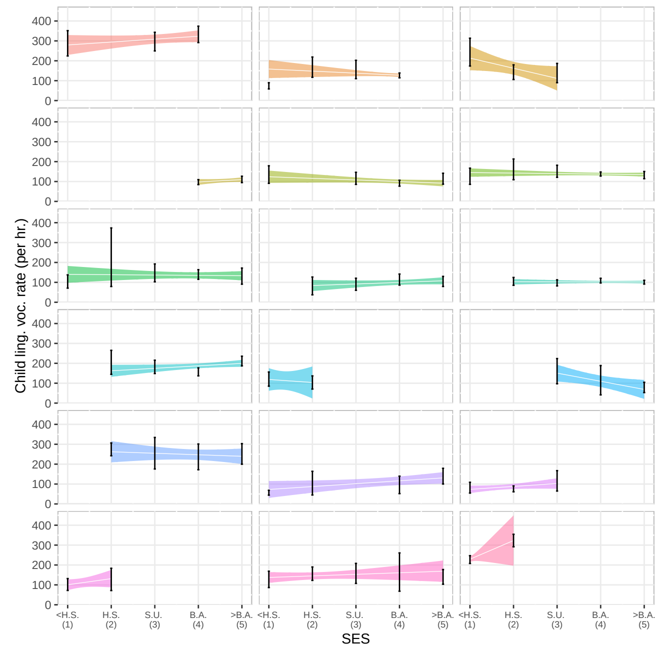

Fig. 4. Child speech as a function of SES within individual corpora. SES = maternal education levels as in Table 1. White lines = linear fit with 95% CIs in color, color = corpus. Black lines = 99% CIs of sample means bootstrapped separately from linear fit for each level of SES. These data (as well as our main models and further analyses in SI 3H/G) do not reveal an SES effect on child speech.

see that age and normative status were significant predictors of adult talk. Instead, we find that neither these (nor any other variables in our model) predicted the quantity of adult talk (Figure 3G). Nonetheless, the precise statistical analyses we carried out do not allow us to directly rule out any of the explanations, a combination of which may be jointly true. Establishing a precise causal chain will require careful consideration of a variety of proximal and ultimate pathways through which child and adult behaviors are shaped. As one example, given that most children here are genetically related to their adult caregivers, we may be observing *covariance* in amount of talk and its linguistic correlates (Explanation 3). Evaluating these alternatives requires evidence from children raised by unrelated caregivers or from genome-wide association studies, as genetic and environmental factors remain challenging to disentangle (27). In this vein, recent work with adopted 15–73-month-olds provides evidence for input effects (maternal utterance length and/or lexical diversity) on adopted children's vocabulary size (measured via caretaker checklist) (28). This study suggests that shared genetics is not the sole contributor to links between (at least these proxies for) caretaker input and child language outcomes. Moreover, shared genetics is just one of the ways in which adult and child behavior may be independently shaped by an unmeasured confounded variable (as per Explanation 3). For instance, other third variables related to dimensions like personality, neighborhood, and childcare context too may be contributors (29, 30). These explanations can only be definitively teased apart by future work.

New Insight on Child and Family Factors. Our main models, figures showing the raw data, and additional analyses (in the North American subset of the data, as well as using an alternative algorithm, see SI3F) reveal effects of normativity, age, and adult talk but not SES (measured here through maternal education), child gender, or monolingualism. To illustrate the complexities involved in determining causal links between child and family factors and child language skills, we again consider how causal links might manifest, using SES as a central example.

Our findings bear on debates regarding SES-associated academic achievement differences in Western industrialized societies (31, 32). Slower language development has often been attributed to parents from lower-SES backgrounds providing less input to their children (viewed from a middle-class Western-centric perspective (32)), leading to calls for behavioral interventions aiming to increase it. Proponents of such interventions might highlight our correlation between adult talk and child speech; critics might instead underscore our finding that SES was not significant in our main analyses nor in every other re-analysis we attempted (SI3E:G).

A full understanding of how SES may relate to children's language input is complicated for empirical and conceptual reasons, leaving strong conclusions premature. On the empirical side, two recent meta-analyses have investigated SES–input correlations, one focused on LENA™ measures (15), and the other based on human-annotated measures (mostly from short lab recordings) (16). The former finds evidence consistent with a publication bias; correcting this bias statistically nearly halves the association between SES and LENA™'s adult talk measure (r= .19 versus .12). The latter finds a sizeable SES effect when inspecting infant-directed speech (r = .34) and a much smaller one when analyzing overall input quantities (r = .09). Together, these studies suggest that our best estimate of the association between overall input quantities and SES is small (r = .1) and may not be detectable even with a sample as large as ours (where the effect was estimated at |d| = .06, or |r| = .03, which did not reach the threshold for significance). Similarly, descriptive plots of the potential correlation between our SES proxy and children's speech (Figure 4) did not suggest a strong or stable relationship across the 18 corpora, leading to our conclusion that, in the sample as a whole, on average, maternal education does not predict how much adults and children talk.

On the conceptual side, SES differences in input and language skills may depend on how language is measured (33). For instance, we speculate that SES effects may be magnified by measures like prevalence of low-frequency words and complex sentence structures common in written text. Such words and structures may occur more in the input to Western, higher-SES children because of parenting practices stereotypical in these groups (34). Moreover, such measures may predict academic achievement better than others, because of the importance literacy has in Western schooling today. In contrast, SES differences in input may be minimized by holistic measures of speech quantities. Indeed, a strength of daylong recordings is that they provide a relatively neutral (rather than Western, high SES-centric) measure, as they tap into how much children are contributing (via speech) to their community's conversational interactions instead of how many rare words or complex constructions they have been taught.

An exclusive focus on word counts or speech quantities likely misses certain behaviors. As machine learning advances (35), it may soon be possible to automatically transcribe conversations happening in daylong recordings (at least in monolingual high-resource language contexts). We suspect that analysis of conversational content may reveal SES differences in, e.g., rare word use or family practices around book-reading even in naturalistic samples (36). Future work with a high-density longitudinal lens is also needed to assess the predictive value of global quantitative measures of speech (like those we employ) relative to more specialized measures (e.g. book-reading practices) with respect to culturally-relevant outcomes (e.g. academic achievement, pragmatic competence in multi-party conversation, etc.)

In our view, causal links between parental behavior and children's outcomes can best be illuminated by randomized control trials. Discovering and leveraging such links to change long-term language outcomes depends on community partnership-based approaches that are informed by the role that structural inequalities play in these outcomes and engage with culturally informed perspectives (37). The present results should not be used to deny families access to resources that evidence suggests are linked with better outcomes for children and their families.

Complicated causal effects are integral to all developmental processes. While we illustrated this with our SES null results, we also found no differences in child speech or adult talk as a function of child gender or multilingual status. Regarding multilingualism, we could not examine relative input in each language the child was exposed to. Future machine learning advances will permit the separate quantification of different languages in daylong recordings, but this must happen alongside reflection on how to fairly measure input and outcomes in such heterogeneous populations (38–40).

Automated Tools and What They Count. A key benefit of our approach is that we were able to pool and identically process 40,933 hours of independently-collected data (SI3A). Moreover, unlike parental surveys, clinical assessments, lab instruments, or hand-annotated data, current published evidence suggests that the LENA™ algorithm's results do not vary systematically by language (though they do vary somewhat across samples, 12). More relevant here, in analyzing the algorithm's accuracy as a function of samples grouped by language and cultural features, we found no significant differences (Methods, SI1E).

While children's language skills grow dramatically over 2–48 months, our measure is not an index of comprehension (which can show quite a different trajectory, 41) but rather of observable linguistic behavior, focusing exclusively on children's rate of linguistic vocalizations (SI3B). These results certainly do not deny effects found on proxies of more narrow-scoped linguistic developments (e.g. vocabulary, processing efficiency, or syntactic complexity), given that some predictors that fail to explain variance here may nonetheless be significant there (3, 42).

The same holds for our measure of adult talk, which is quantitative and holistic; additional research is needed to distinguish child-*directed* from child-*available* speech, with the latter including all speech the child hears. Although some research suggests child-directed speech shows tighter correlations with children's vocabulary than child-available speech does (43, 44), the importance of the latter has not been as fully studied for other types of language knowledge (45); and, as far as we know, this paper is the first to document a significant link for everyday child speech behavior. Therefore, it would be relevant to further investigate the strength of the predictive value of overall adult talk (which was a significant predictor here) versus child-directed talk, in a similarly large and diverse sample as the present one. Unfortunately, automated tools for separating child-directed from overheard speech are not yet sufficiently accurate to make this possible (46). Future work could also develop promising new approaches for considering other sources of speech (e.g., other children) given their relevance as a function of family structure (47). These approaches were not possible here due to both technical algorithmic constraints and family structure information not being available in our data-subsets. Another fruitful future direction could consider conversational dynamics, studying both children's tendency to vocalize around adults and the complexity of such vocalizations. Recent work (that is critically reliant on human annotation of social intent) raises particularly interesting ideas in this domain (14, 48). Relatedly, novel exploratory analyses describing the acoustics of children's vocalizations (49) hold promise for driving future hypothesis-testing work building on the present results.

Whatever measures are employed in the future as proxies of child language production and input, we strongly encourage researchers to consider psychometric properties and ecological validity. The current approach demonstrates measure validity that is comparable to that of other standard infant instruments (SI1D:E). As context, measures used as proxies for infant language and cognitive knowledge are inherently noisier than the best batteries used to assess highly educated adults in

Western-centric settings. Notably, even there, reliabilities can fall well below r = 1.[‖]

Moreover, standardized tests face ecological validity threats, particularly when applied cross-culturally. If our goal is to measure and understand the human mind, we need implementable, culturally sensitive and appropriate ways of measuring human behavior on a large scale. To our knowledge, there are no such measures whose reliability has been examined, driving us to conduct extensive quantification of the reliability of the metrics we employed here (SI1D:E). We found that our measures show levels of reliability that are consistent with those already in use for research and clinical purposes in infant populations. For example, the MacArthur-Bates Communicative Development Inventory (a parental report instrument used largely as a proxy for vocabulary size) has been the basis for cross-linguistic, demographic, and clinical research (9, 51–53), and reports a median correlation between itself and laboratory measures of .61 (54). Our median accuracy comparing automated and manual annotation for each of our algorithms (LENA™ and VTC) is .74, squarely in line with field standards (SI1E). Indeed, converging evidence across these two wholly separate algorithms regarding overall accuracy of our measure serves to increase confidence in the validity of our results.

In sum, rather than eliciting knowledge or caregiver-child interaction in a constrained lab setting, or using checklists in contexts where they make little sense socio-culturally, we measure everyday language use *en masse*. Our measure of early speech production is global, since we simply measure more versus less speech or speech-like production on the part of adults and children as they go about their daily life. And yet, these measures have important advantages, which led us to select them as proxies here, including comparable reliability to other measures of language development commonly used in both research and applied settings (Methods, SI1D:E); reported correlations between them and finer-grained, "qualitative" measures of language development (SI1D), and convergent validity with respect to standardized language tests (13). Most importantly, our speech measure merits consideration as one of many possible proxies of language development thanks to its cross-cultural adaptability, observer-free sampling volume, and sheer ecological validity. Indeed, our results raise the possibility that more ecologically-valid lexical, phonetic, or grammatical measures will also reveal stability across factors like SES (55), gender, and multilingualism. Exploring these factors, however, awaits machine-learning developments that can extract such fine-grained linguistic measures from the raw audio collected with child-worn devices.

Conclusion. Our analysis of speech behavior in daily life around the world evinces scientific progress on two fronts. First, by revealing substantial variation in young children's speech, we provide evidence against a monolithic picture of language development. Instead, this work reveals individual variation as *fundamental* to our understanding of this species-wide ability. Second, by tapping into natural speech interactions at unprecedented scale and diversity, we are able to move beyond prior work by simultaneously considering the interlocking factors that affect speech production over early development. Our results reveal not only the expected correlations with age and clinical factors, but also substantial associations with adult talk. All other factors paled in comparison with these three, the null effect of our SES proxy being of particular noteworthiness. These findings open exciting avenues for both theoretical research and potential applications, including the prospect of behavioral interventions to harness adult talk in the context of speech and language diagnoses. Small-scale experimental and observational research has been fundamental to our understanding of language, development, and the human mind. Machine learning (like that in speech technology) promises to extend our scientific reach by exploding the range of everyday interactions we are able to capture and analyze. Just as recent technological innovations have opened new vistas in understanding the vocalizations of mice and whales (6, 7), so too does speech technology have the potential to reveal how everyday human communication gives rise to language learning in children around the world.

Methods

All code used to generate our analysis and the manuscript is available at https://osf.io/9v2m5/?view_only=50df17fcf0844145ae692c35b78c6b08.

Data Discovery and Integration. We took steps to counter a prevalent bias for normative North American data (see SI3A for corpus constitution procedure). Included data were independently collected by 18 stewards (56–77); see SI5 for list of publications based on individual datasets. We note that while our corpora covered a much greater variety of participants than prior work, it would not be appropriate to interpret our samples as comprehensively representative of the country or language community from which they are drawn.

Socioeconomic status and normative development were streamlined for cross-corpus consistency (SI2A:B, SI3A, Figure S3A.1). For socioeconomic status we use maternal education, a reliable proxy for SES in previous research on language development (18, 78). Maternal education was available across all datasets, and could be converted into a 5-point maternal education scale with levels corresponding to less than high school degree, high school degree or equivalent, some college/vocational/associate degree level training, university/college degree, and advanced degree (SI2B; Table S2B.1).

For non-normative development, data stewards had tagged a wide variety of infant or familial characteristics as potentially non-normative. We confirmed that the classification was backed up by extant literature (SI2A). Infants ultimately classified as having non-normative development in the present sample include those who met one or more of the following criteria: preterm birth (<37 weeks); diagnosed speech or language delay; global developmental delay; low birth weight (<2500g when specified); hearing loss, hearing aids or cochlear implants; familial risk of Autism Spectrum Disorder, specific language impairment and/or dyslexia; other relevant genetic syndromes. Notably, our child vocalization rate measure is not a standardized normed clinical evaluation, and thus non-normative status may not necessarily translate to behavior that falls >1 standard deviations below the norm in these naturalistic recordings.

Analysis Details. We first randomly partitioned the data within each corpus such that 35% of monolingual, normative children were placed in an exploration set (N children = 264; N

[‖] For instance, prior work finds test-retest reliabilities as low as r = .6 for certain sections of the widely used Wechsler Adult Intelligence Scale among North American English-speaking adults (50).

recordings = 850), and all others in a confirmation set (N children = 737; N recordings = 2025) (SI3A). The exploration set was used to study the psychometric properties of potential language input and output variables (SI3B), resulting in the selection of the output variable referred to as **child speech** above, and CVCr (Child Vocalization Count rate) in analysis and supplementary files (SI3B, Table S3B.1); and the input variable referred to as **adult talk** above, and AVCr (Adult Vocalization Count rate) in analysis and supplementary files (SI3B, Table S3B.2). Note that this includes both child-directed and child-available speech.

In addition, we used the exploration set to check the robustness of results to variation in random effect structure, and explored diverse model structures using mixed models in R's lme4 package (79), checking whether the addition of effects or interactions explained additional variance (SI3C). This led us to (a) include overlap rate as a covariate (see Figure S3C.1), to control for the fact that in noisy environments, more child speech and adult talk within the same recordings may be labeled as "overlap" by LENA (and thus not attributed to either speaker type) and (b) to not include random slopes for any of the predictors. Regarding the latter choice, our exploration of random effect structure revealed that models including random slopes for any of the predictors (notably including gender and SES) as a function of corpus led to non-convergent models. While such non-convergence could be due to various reasons, the most likely explanation is that the model is overparametrized (80), i.e., variance cannot be reliably attributed to predictors *within* each corpus (see SI3H for additional checks, including one including random slopes for SES, and SI2B for discussion of alternatives to our SES implementation).

Evaluation against human annotations. To assess the validity of our child speech and adult talk measures, we evaluated them against human annotations (see SI1D:E for further information). The median correlation of human to algorithm performance for the algorithms is >.7, i.e. comparable reliability to established developmental clinical and research instruments (81–83). As far as we know, the present multi-cultural validation exceeds those from prior research instruments. For example, the Ages and Stages Questionnaire (84) is a standard instrument used at well-child visits in the U.S. It is also recommended by the World Bank as one of the most popular tools to measure child development, used in at least 20 countries (85). And yet, a recent systematic review (83) reports only 6 reliability analyses (averaging, e.g., .7 for internal consistency at 24mo.). Relative to this, our validation effort containing estimates for 14/18 corpora and finding strong validity is notable. Finally, one may wonder whether the LENA™ algorithm performs less well for languages and cultures that diverge from its training set, which was English-learning children growing up in an urban/suburban U.S. setting. Although we observe considerable corpus variation, this variation is not attributable to whether children were learning English or growing up in an urban setting, as assessed by Welch's t-tests, for either our child speech measure (CVCr; English versus non-English medians 0.785 vs. 0.71, $t(6.04)$ = -0.5, $p = 0.637$; urban versus rural medians 0.77 vs. 0.71, $t(8.11)$ = -0.46, $p = 0.661$), or for our adult talk measure (AVCr; English versus non-English medians 0.75 vs. 0.74, $t(7.91) = 0.42$, $p = 0.686$; urban versus rural medians 0.75 vs. 0.74, $t(3.07)$ = -0.23, $p = 0.835$). Instead, our results suggest that corpus variation more likely reflects how the human annotation was done rather than how well the algorithm worked, since the corpora with lower reliabilities were also those in which the human annotation was more coarse-grained (see SI1E).

Additional algorithm. To make sure that key conclusions were robust to methodological details, we reanalyzed the subset of the data for which data stewards shared audio with a newer, open-source alternative to LENA™: the Voice Type Classifier (VTC) (86). Like the LENA™ algorithm, VTC returns an estimation of child and adult vocalization counts. A total of 1065 audio files from 11 corpora were available for this reanalysis (SI3F).

The VTC algorithm employs a completely different approach than the proprietary algorithm developed by LENA™, including the use of neural networks running directly from the audio (rather than from MFCC features). VTC allows multiple talker classes to be activated at the same time, whereas in the LENA™ algorithm, overlap between talkers (or between a talker and noise) is tagged as "Overlap," which is not counted towards children's input or output. VTC also differs from LENA™ in its training set. While LENA™ was trained entirely on data from North American, monolingual English-learning, urban children, VTC was developed using the combination of various corpora of children residing in urban or rural settings and learning one or more of several languages (including the tonal language Minn, French, Ju|'hoan, Tsimane, English, and several others, in rough order of quantity of data). Further information on accuracy is provided in SI1E; both algorithms render similar accuracy when compared to human annotation as noted above.

Models. We used linear mixed regressions (Gaussian family), and established model structure from the exploration data (SI3C). Hypotheses were derived from exploratory models and systematic reviews of literature on monolingualism and normativity (SI3D). The model predicting the rate of children's linguistic vocalizations (i.e. child speech) was: $child\_gender + SES + child\_normative * AVCr * age + child\_monolingual * AVCr * age + overlap + (1 + overlap + AVCr|corpus) + (1|corpus : child\_id)$. The model predicting the rate of adult linguistic vocalizations (i.e. adult talk) was: $child\_gender + SES + child\_normative * age + child\_monolingual * age + overlap + (1 + overlap|corpus) + (1|corpus : child\_id)$. Full model details and a link to model diagnostics are provided in SI3E. We report estimates (standardized, which serve as effect sizes), standard errors of the estimates, and q-values (FDR-corrected p-values); see Tables 1 and 2.

Participants. Table 3 lists participant characteristics noting both (1) the exploration/confirmation split (SI3A), and (2) that some children provided multiple recordings. We excluded 2/850 recordings from 1/264 children from the exploration set and 8/2025 recordings from 5/737 children in the confirmation set from our models because data regarding their maternal education was missing. For child gender, there were slightly more boys than girls. This was in part because corpora with children with non-normative development also include children with normative development matched in gender, leading to an over-representation of boys since more boys than girls have non-normative development. See Table 3 and Figure 5 for specific numbers and visualized distributions.

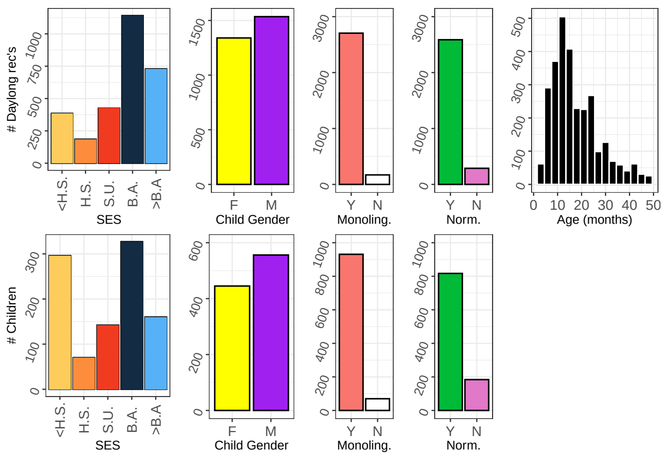

Fig. 5. Sample demographics. Number of daylong recordings (top row) and children (bottom row) in the full dataset across demographic variables. For socioeconomic status (SES), <H.S. = less than high school degree, H.S. = high school degree, S.U. = some university, B.A. = bachelor's degree, >B.A. = advanced degree. For child gender, F = female, M = male. For monolingual status (monoling.), Y = monolingual, N = not monolingual. For normative development (norm.), Y = normative, N = non-normative.

Language Background. The languages represented in these data covered many languages and language families. Using classifications from Glottolog (87), we report that our 18 corpora feature 10 primary languages (Dutch, English, Finnish, French, Spanish, Swedish, Tsimane, Vietnamese, Wolof, Yélî Dnye) from 5 distinct language families and one isolate (Atlantic-Congo, Austroasiatic, Indo-European, Mosetén-Chimané, Uralic, Yélî-isolate); see Figure 1. Based on corpus metadata provided by each data steward, the recorded children were also exposed to an additional 33 languages (Arabic, ASL, Berber, Cantonese, Croatian, Danish, Farsi, Frisian, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Khmer, Korean, Macedonian, Malay, Malayalam, Mandarin, Norwegian, Papiamento, Polish, Portuguese, Romanian, Russian, Sahaptin, Slovenian, Solomon-Islands Pidgin, Thai, Turkish, Yoruba), which add 11 further language families (Afro-Asiatic, Austroasiatic, Austronesian, Deaf Sign Languages—LSFic, Dravidian, Japonic, Koreanic, Sahaptian, Sino-Tibetan, Tai-Kadai, Turkic) and bolster data from three language families already represented by the primary languages (Atlantic-Congo, Indo-European, and Uralic).

ACKNOWLEDGMENTS. We thank Adriana Weisleder, Ann Weber, Camila Scaff, Karmen McDivitt, Evan Kidd, Bridgette Kelleher, Hillary Ganek, Anne Fernald, Hanna Elo, Samantha Durrant, Yatma Diop, John Bunce, and Sarp Uner for organizing and/or sharing their data with us. The authors acknowledge the following funding sources: ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017 (AC); J. S. McDonnell Foundation (AC); NEH HJ-253479-17 (EB); NIH DP5-OD019812 (EB); NSF BCS-1844710 (EB), NSF SBE-0354453 (NRE); ESRC ES/L008955/1 (CR); SSHRC 435-2015-0628, 869-2016-0003 (MS); NSERC 501769-2016-RGPDD (MS); Netherlands Organisation for Scientific Research 275-89-033 (MC); NIMH K23MH111955; NIDCDD F31DC018219 (LH); MAW 2011.0070 (ICS, EM); MAW 2013.0056 (ICS, EM); Marie Sklodowska-Curie Individual Fellowships European Program 798908 (MK); ARC CE140100041 (Evan Kidd).

1. Pinker S (1994) *The language instinct* (Morrow, New York).
2. Oller DK, et al. (2020) Infant boys are more vocal than infant girls. *Current Biology* 30(10):R426–R427.

Table 3. Number of children and recordings by demographic variables, split by exploration and confirmation subsets.

		Exploration Subset		Confirmation Subset	
Variables	Levels	Children	Recs.	Children	Recs.
Gender	Boys	156	516	398	1016
	Girls	107	332	334	1001
Normativity	Normative	263	848	550	1731
	Non-normative	0	0	182	286
Lingualism	Monolingual	263	848	662	1847
	Multilingual	0	0	70	170
SES	<H.S. (1)	94	120	202	265
	H.S. (2)	10	26	60	159
	S.U. (3)	27	116	115	309
	B.A. (4)	86	355	241	786
	>B.A. (5)	46	231	114	498
Total N		263	848	732	2017

Note. Children = # of children; Recs. = # of daylong recordings. In SES, <H.S. = children whose mothers have (the equivalent of) less than a high school degree; H.S. = high school degree; S.U. = some university; B.A. = bachelor's degree; >B.A. = more than a bachelor's degree. Multilingual children, children with non-normative development, and 65% of all other children were reserved for the confirmation subset. N.B. the 6 children with missing data for maternal education are omitted from this table.

3. Fernald A, Marchman VA, Weisleder A (2013) SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science* 16(2):234–248.
4. Gilkerson J, et al. (2017) Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology* 26(2):248–265.
5. Coe R (2002) It's the Effect Size, Stupid: What effect size is and why it is important. Available at: https://f.hubspotusercontent30.net/hubfs/5191137/attachments/ebe/ESguide.pdf [Accessed July 28, 2021].
6. Coffey KR, Marx RG, Neumaier JF (2019) DeepSqueak: A deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology* 44(5):859–868.
7. Shiu Y, et al. (2020) Deep neural networks for automated detection of marine mammal species. *Scientific Reports* 10(1):607.
8. Broesch T, et al. (2020) Navigating cross-cultural research: Methodological and ethical considerations. *Proceedings of the Royal Society B: Biological Sciences* 287(1935):20201245.
9. Frank MC, Braginsky M, Marchman VA, Yurovsky D (2021) *Variability and Consistency in Early Language Learning: The Wordbank Project.* (MIT Press, Cambridge, MA) Available at: https://langcog.github.io/wordbank-book/index.html#.
10. Zimmerman FJ, et al. (2009) Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics* 124(1):342–349.
11. Bergelson E, Amatuni A, Dailey S, Koorathota S, Tor S (2019) Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science* 22(1):e12715.
12. 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.

13. Wang Y, Williams R, Dilley L, Houston DM (2020) A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review* 57:100921.
14. Oller DK, et al. (2019) Preterm and full term infant vocalization and the origin of language. *Scientific Reports* 9(1):14734.
15. Piot L, Havron N, Cristia A (2022) Socioeconomic status correlates with measures of Language Environment Analysis (LENA) system: A meta-analysis. *Journal of Child Language* 49(5):1037–1051.
16. Dailey S, Bergelson E (2022) Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. *Developmental Science* 25(3):e13192.
17. Richman AL, Miller PM, LeVine RA (1992) Cultural and educational variations in maternal responsiveness. *Developmental Psychology* 28:614–621.
18. Hoff E (2003) The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development* 74(5):1368–1378.
19. Hartas D (2011) Families' social backgrounds matter: Socio-economic factors, home learning and young children's language, literacy and social outcomes. *British Educational Research Journal* 37(6):893–914.
20. Rowland CF, Alcock K, Meints K (2022) The (null) effect of socio-economic status on the language and gestures of young infants: Evidence from British English and eight other languages Available at: https://osf.io/hwg4c [Accessed April 21, 2023].
21. Hackman DA, Farah MJ (2009) Socioeconomic status and the developing brain. *Trends in cognitive sciences* 13(2):65–73.
22. UNESCO Institute for Statistics (2012) *International Standard Classification of Education (ISCED) 2011* (UNESCO Institute for Statistics) doi:10.15220/978-92-9189-123-8-en.
23. Oller DK, et al. (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.
24. Rankine J, et al. (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.
25. McDaniel J, et al. (2020) Effects of pivotal response treatment on reciprocal vocal contingency in a randomized controlled trial of children with autism spectrum disorder. *Autism*:1362361320903138.
26. Kidd E, Donnelly S (2020) Individual Differences in First Language Acquisition. *Annual Review of Linguistics* 6(1):319–340.
27. Bishop DVM (2014) Ten questions about terminology for children with unexplained language problems. *International Journal of Language & Communication Disorders* 49(4):381–415.
28. Coffey JR, Shafto CL, Geren JC, Snedeker J (2022) The effects of maternal input on language in the absence of genetic confounds: Vocabulary development in internationally adopted children. *Child Development* 93(1):237–253.
29. Hilton M, Twomey KE, Westermann G (2019) Taking their eye off the ball: How shyness affects children's attention during word learning. *Journal of Experimental Child Psychology* 183:134–145.
30. De Marco A, Vernon-Feagans L (2013) Rural Neighborhood Context, Child Care Quality, and Relationship to Early Language Development. *Early Education and Development* 24(6):792–812.
31. Golinkoff RM, Hoff E, Rowe ML, Tamis-LeMonda CS, Hirsh-Pasek K (2019) Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development* 90(3):985–992.
32. Sperry DE, Sperry LL, Miller PJ (2019) Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development* 90(4):1303–1318.
33. Ochs E, Kremer-Sadl T (2020) Ethical Blind Spots in Ethnographic and Developmental Approaches to the Language Gap Debate: *Langage et société* N° 170(2):39–67.
34. Dickinson DK, Griffith JA, Golinkoff RM, Hirsh-Pasek K (2012) How Reading Books Fosters Language Development around the World. *Child Development Research* 2012:e602807.
35. Lavechin M, et al. (2022) Brouhaha: Multi-task training for voice activity detection, speech-to-noise ratio, and c50 room acoustics estimation. *arXiv preprint arXiv:221013248*.
36. Nutbrown C, et al. (2016) Families' roles in children's literacy in the UK throughout the 20th Century. *Journal of Early Childhood Literacy* 17. doi:10.1177/1468798416645385.
37. Weber A, Fernald A, Diop Y (2017) When Cultural Norms Discourage Talking to Babies: Effectiveness of a Parenting Program in Rural Senegal. *Child Development* 88(5):1513–1526.
38. Bialystok E, Werker JF (2017) Special issue: Systematic effects of bilingualism on children's development. *Developmental Science* 20(1):e12535.
39. Oller DK, Pearson BZ, Cobo-Lewis AB (2007) Profile effects in early bilingual language and literacy. *Applied Psycholinguistics* 28(2):191–230.
40. Grüter T, Hurtado N, Marchman VA, Fernald A (2014) Language exposure and online processing efficiency in bilingual development. *Input and Experience in Bilingual Development* (John Benjamins Publishing Company), pp 15–36.
41. Clark EV, Hecht BF (1983) Comprehension, Production, and Language Acquisition. *Annual Review of Psychology* 34(1):325–349.
42. Eriksson M, et al. (2012) Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology* 30(2):326–343.
43. Shneidman LA, Arroyo ME, Levine SC, Goldin-Meadow S (2013) What counts as effective input for word learning? *Journal of Child Language* 40:672–686.
44. Weisleder A, Fernald A (2013) Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science* 24(11):2143–2152.
45. Cristia A (2020) Language input and outcome variation as a test of theory plausibility: The case of early phonological acquisition. *Developmental Review* 57:100914.
46. Schuller B, et al. (2017) The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring. *Interspeech 2017* (ISCA), pp 3442–3446.
47. Cristia A, Gautheron L, Colleran H (2023) Vocal input and output among infants in a multilingual context: Evidence from long-form recordings in Vanuatu. *Developmental Science* n/a(n/a):e13375.
48. Pretzer GM, Lopez LD, Walle EA, Warlaumont AS (2019) Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe. *Infant Behavior and Development* 57:101325.
49. Ritwika VPS, et al. (2020) Exploratory dynamics of vocal foraging during infant-caregiver communication. *Scientific Reports* 10(1):10469.
50. Strauss E, et al. (2006) *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (Oxford University Press).

51. Thal DJ, Bates E, Goodman J, Jahn-Samilo J (1997) Continuity of language abilities: An exploratory study of late- and early-talking toddlers. *Developmental Neuropsychology* 13(3):239–273.
52. Thal DJ, O'Hanlon L, Clemmons M, Fralin L (1999) Validity of a Parent Report Measure of Vocabulary and Syntax for Preschool Children With Language Impairment. *Journal of Speech, Language, and Hearing Research* 42(2):482–496.
53. Thal D, DesJardin JL, Eisenberg LS (2007) Validity of the MacArthur–Bates Communicative Development Inventories for Measuring Language Abilities in Children With Cochlear Implants. *American Journal of Speech-Language Pathology* 16(1):54–64.
54. Fenson L, et al. (1994) Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development* 59(5):i–185.
55. Villar J, et al. (2019) Neurodevelopmental milestones and associated behaviours are similar among healthy children across diverse geographical locations. *Nature Communications* 10(1):1–10.
56. Bergelson E (2017) *Bergelson Seedlings HomeBank Corpus* Available at: doi:10.21415/T5PK6D.
57. Brookman R, et al. (2020) Mother-infant interactions and expressive language development: The effects of maternal depression and anxiety. *Child Development*.
58. Canault M, Le Normand M-T, Foudil S, Loundon N, Thai-Van H (2016) Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods* 48(3):1109–1124.
59. Cristia A, Casillas M (2020) LENA recordings gathered from children growing up in Rossel Island.
60. Elo H (2016) Acquiring Language as a Twin: Twin children's early health, social environment and emerging language skills. PhD thesis (Tampere University). Available at: http://urn.fi/URN:ISBN:978-952-03-0296-2.
61. Ganek H, Eriks-Brophy A (2019) *LENA its data from daylong recordings collected in Vietnam* Available at: osf.io/d9453.
62. Hamrick L, Seidl A, Tonnsen BL (2019) *LENA its data from daylong recordings gathered from children with typical and atypical development* Available at: osf.io/n9pvq/.
63. Kidd E, Junge C, Spokes T, Morrison L, Cutler A (2018) Individual Differences in Infant Speech Segmentation: Achieving the Lexical Shift. *Infancy* 23(6):770–794.
64. Marklund E, Schwarz I-C, Lacerda F (2020) *LENA its-data from daylong recordings in Swedish-speaking families with 3- to 10-month-olds (recorded 2016)* Available at: osf.io/wh9dt.
65. McDivitt K, Soderstrom M (2016) *McDivitt HomeBank Corpus* Available at: 10.21415/T5KK6G.
66. Ramírez-Esparza N, García-Sierra A, Kuhl PK (2014) Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science* 17(6):880–891.
67. Ramírez-Esparza N, García-Sierra A, Kuhl PK (2017) The impact of early social interactions on later language development in Spanish–English bilingual infants. *Child Development* 88(4):1216–1234.
68. Rowland CF, Bidgood A, Durrant S, Peter M, Pine JM (2017) *The Language 0–5 Project Corpus* Available at: https://nyu.databrary.org/volume/389.
69. Scaff C, Stieglitz J, Cristia A (2020) *Tsimane' daylong recordings collected with LENA in 2017-2018* Available at: DOI 10.17605/OSF.IO/6NEZA.
70. Schwarz I-C, Marklund E, Gerholm T (2019) *LENA its-data from daylong recordings in Swedish-speaking families with 30-month-olds (recorded 2016)* Available at: osf.io/yzp4b.
71. Schwarz I-C, Marklund E, Lam-Cassettari C, Marklund U (2019) *Longitudinal LENA its-data from daylong recordings in Swedish-speaking families with infants at 6, 12, 16 and 24 months.*
72. Van Alphen P, Meester M, Dirks E (2020) *LENA onder de loep; ITS files and metadata of daylong LENA recordings at the homes of preschoolers with DLD and TD peers (collected by the Royal Dutch Kentalis and the NSDSK)* Available at: osf.io/2zyub.
73. VanDam M (2018) *VanDam Public 5-minute HomeBank Corpus* Available at: doi:10.21415/T5388S.
74. Warlaumont AS, Pretzer GM, Mendoza S, Walle EA (2016) *Warlaumont HomeBank Corpus* Available at: doi:10.21415/T54S3C.
75. Weber A, Marchman VA, Fernald A (2019) *LENA its data collected in Kaolack Senegal in 2013* Available at: https://doi.org/10.17605/OSF.IO/EMBFS.
76. Weisleder A, Mendelsohn A (2019) *Daylong recordings of 2-12 month-old infants from Spanish-speaking homes in the US* Available at: DOI 10.17605/OSF.IO/JBTNC.
77. Van Alphen P, Davids N, Dijkstra E, Fikkert P (2020) *TiBLENA: ITS files and metadata of daylong LENA recordings at the homes of preschoolers with DLD and TD peers (collected by the Royal Dutch Kentalis and the Radboud University)* Available at: osf.io/ymv7b.
78. Bornstein MH, Hahn C-S, Suwalsky JTD, Haynes OM (2003) Socioeconomic status, parenting, and child development: The Hollingshead Four-Factor Index of Social Status and The Socioeconomic Index of Occupations. *Socioeconomic Status, Parenting, and Child Development*, Monographs in parenting series. (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US), pp 29–82.
79. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1):1–48.
80. Bates D, Kliegl R, Vasishth S, Baayen H (2018) Parsimonious Mixed Models. *arXiv:150604967 [stat]*. Available at: http://arxiv.org/abs/1506.04967 [Accessed April 29, 2022].
81. Dale PS (1991) The Validity of a Parent Report Measure of Vocabulary and Syntax at 24 Months. *Journal of Speech, Language, and Hearing Research* 34(3):565–571.
82. Feldman HM, et al. (2005) Concurrent and Predictive Validity of Parent Reports of Child Language at Ages 2 and 3 Years. *Child Development* 76(4):856–868.
83. Velikonja T, et al. (2017) The psychometric properties of the Ages & Stages Questionnaires for ages 2-2.5: A systematic review. *Child: Care, Health and Development* 43(1):1–17.
84. Bricker D, et al. (1999) Ages and stages questionnaire. *Baltimore, MD: Paul H Brookes.*
85. Fernald LCH, Prado E, Kariger P, Raikes A (2017) A Toolkit for Measuring Early Childhood Development in Low and Middle-Income Countries. *MINISTERIO DE EDUCACIÓN*. Available at: https://repositorio.minedu.gob.pe/handle/20.500.12799/5723 [Accessed May 11, 2022].
86. Lavechin M, Bousbib R, Bredin H, Dupoux E, Cristia A (2020) An open-source voice type classifier for child-centered daylong recordings. *Interspeech*. Available at: http://arxiv.org/abs/2005.12656 [Accessed September 11, 2020].
87. Hammarström H, Forkel R, Haspelmath M, Bank S (2020) *Glottolog 4.2.1* (Max Planck Institute for the Science of Human History, Jena) Available at: https://glottolog.org/ [Accessed June 4, 2020].