Interspeech 2024
1-5 September 2024, Kos, Greece

The Difficulty and Importance of Estimating the Lower and Upper Bounds of Infant Speech Exposure

Joseph Coffey[1], *Okko Räsänen*[2], *Camila Scaff*[1,3], *Alejandrina Cristia*[1]

[1] PSL University, Laboratoire de Sciences Cognitives et de Psycholinguistique (ENS, EHESS, CNRS, DEC), France
[2] Tampere University, Unit of Computing Sciences, Finland
[3] University of Zurich, Institute of Evolutionary Medicine (IEM), Switzerland

jrcoffey@g.harvard.edu, okko.rasanen@tuni.fi, camillescaff@gmail.com, alecristia@gmail.com

Abstract

Estimates of infants' language exposure are necessary for computational studies that attempt to model and learn from infant language experiences. However, there are no well-established input estimates usable for this purpose. This paper explores empirical data on infant language exposure across various cultural settings to derive plausible limits on the speech exposure infants might receive during their first years of life. First, we argue that several assumptions lack unanimous agreement and demonstrate that existing data are problematic in multiple ways. Integrating these uncertainties and published information, we find estimates that range from 1 to 3,300 hours per year. We end by discussing how such a large possible range may impact evaluation of the plausibility and benchmarking of computational models.

Index Terms: language acquisition, computational modeling

1. Introduction

With the rise of deep learning and self-supervision, models that process speech similarly to humans seem within reach [1, 2]. These models take raw auditory and/or multimodal (e.g., audiovisual) signals as input. Researchers can then benchmark their outcomes against infants' language acquisition landmarks [3]. The goal of these models is not only to understand how language learning operates in humans, but also to advance the development of more data-efficient machine learning technologies for processing spoken and written language.

Among the most salient remaining questions in this line of work is the plausibility of hypothesized learning mechanisms when paired with ecologically plausible quantities of input. For instance, large language models operating in the written domain achieve human level for many linguistic tasks after training with much more material than even human adults would have experienced in their entire lives[1], but have also recently shown promise on less data [4]. In contrast, models operating on speech audio perform much worse than text-based systems on the same size input [3]. However, the amount of input that would be considered as plausible is not well-defined in the existing research. This raises the question: Where should we set the bar? What is a plausible range of language exposure that, based on human experience, nonetheless results in typical language acquisition?

In this paper, we first lay out the many assumptions that need to be made to draw such estimates. Second, we elaborate on the sources of information one can draw from to try to pinpoint the quantity of input afforded to humans. Third, we explain how assumptions and sources can be crossed to generate estimates. Finally, we discuss the significance of the estimates.

[1] Yann LeCun, https://bit.ly/3V7Rylx

2. Assumptions regarding experience and processing biases

Children are born into a world saturated with language, with speech coming from family, friends, neighbors, radios, phones, and in many countries even televisions sets and children's toys. If we were to base estimates of speech exposure on all potential speech children might hear, we risk greatly overestimating the amount of data children have access to. It is clear from studies of language development that potential speech input is not equivalent to children's perceptual intake. Much of what is spoken around the child does not appear to enter into the learning process for a variety of reasons. While much of the research we base this assertion on is drawn exclusively from the U.S. and Europe [5], this data is sufficient to demonstrate that what constitutes meaningful speech exposure is far more complicated than simply "spoken language."

Even if we restrict our estimates to proximate human-produced speech, attempting to define what input is "meaningful" is still difficult. For instance, there are many studies that suggest children may prefer certain kinds of speech over others. Previous research suggests that infants prefer 'parentese' or child-directed speech (CDS), which features a higher pitch and more exaggerated vowel sounds than adult-directed speech[2]. It has also been found that the presence of parentese is tied to better language outcomes [6], and that child vocabularies are better predicted by CDS than overall quantities both in North-American samples and at least one sample of Yucatec Mayan infants [7, 8]. If only CDS is deemed suitable for language learning, the proportion of relevant speech goes down to somewhere between 14% and 70% of proximate human-produced speech [7, 8, 9].[3]

Moreover, not all speakers may be attended to similarly. For instance, some correlational work suggests children's outcomes can be better predicted from father's input more than mother's input [10]. This is surprising because mothers' speech is much more prevalent, constituting between 30% and 70% of infants'

[2] M. Zettersten et al., preprint available from https://doi.org/10.31234/osf.io/etqs7

[3] Bunce et al., preprint available from https://psyarxiv.com/723pr/download?format=pdf

10.21437/Interspeech.2024-2102

child-directed input[2]. The evidence that input from people other than the parents, including both teachers and other children, count as much as mothers is scarce, and suggest that this may only be in certain situations [11, 7, 12]; for example, in the case of immigrant families when adults do not speak the ambient language, and older siblings do. For the estimations below, we will use as a conservative estimate that 30% of CDS is optimal because it comes from mothers, and a liberal estimate is that 70% does.

Input estimates might also be affected by the nearness and clarity of speech. Children have been found to perform worse than adults on speech recognition tasks presented with background noise [13], which has been linked to worse performance on word-learning tasks [14]. These findings have been supported by environmental studies of children's learning outcomes, which report associations between exposure to ambient noise from nearby traffic and airports at school and impairments in reading and memory tasks [15]. Thus, even speech produced by appropriate speakers may not enter into the learning process. We do not know of estimates in terms of what proportion of children's spoken input overlaps with noise, but the estimated speech-to-noise ratios (SNR) for a North American corpus suggest much of it does: The top decile was 12–24 dB, and the bottom spanned between -9 and -4 dB [16]. Another cross-linguistic study estimated an average SNR of approximately 0 dB for child-centered audio recordings [17]. For our estimations below, we will liberally estimate that 100% of speech input is noise-free, and conservatively estimate that only 30% is.

Among the categories of noise, a particularly challenging one is overlap across speakers, which should make it harder to process because speaker voices will largely overlap in physical characteristics (more than e.g., airport noise does with human voice). In our experience, overlap in natural conversation represents a minute portion of input (as little as 1%), but we can imagine that in certain cultural contexts, it would be much more common than this, with a conservative estimate that 10% of input may be overlapping across speakers.

There are even more basic things we do not know, such as *when* children are exposed to speech and can process it. As we will explain in the next section, many previous studies build on short-term observations, but even when 16 continuous hours are recorded, infants may not be able to process all of that if we assume that the infants need to be awake when they hear speech. More generally, infants may not be equally receptive to speech at different moments throughout the day (e.g., while tired, fussy, or particularly engaged in some activity). Moreover, regardless of whether estimates come from short or long recordings, we need some kind of parameter to extrapolate these estimates to the proportion of a 24-hour period in which the infant is awake. Research shows that this varies as a function of age as well as individuals, with sleeping times ranging between 9 and 14h, which means individual infants may spend between 10 and 15h per day awake [18].[4]

3. Sources of information

There are many potential sources of information with which to estimate how much input infants may be exposed to. For example, we can refer to transcribed samples of input reported in studies of CDS. The largest example of these studies is Hart & Risley's famous "30-million-word gap" study [20], which followed 42 families residing in Kansas over 2.5 years, collecting and transcribing samples of naturalistic speech produced by children and their primary caregivers during monthly hour-long visits. They found that the thirteen children in the highest income group heard 2153 word tokens per hour in CDS from their primary caregiver on average, while the six children in the lowest income group heard 616 words per hour. Across all households, the average rate was 1439 word tokens per hour. Assuming a 14-hour waking day and about 5200 waking hours in a year, over the course of the first four years of life these rates would result in 45-million words of exposure in the highest income group and 15-million words in the lowest income group, resulting in a 30-million-word gap. If we assume roughly .3s per word (as estimated from the Brent corpus for CDS; [21]), this would result in roughly 940 hours per year of CDS in the high income group, 270 hours per year of speech in the low income group, and roughly 620 hours per year on average.

There are, however, some problems with this estimation [9]. Firstly, Hart & Risley represent an outlier in studies of socioeconomic differences in speech exposure, as most studies do not find as pronounced differences between high- and low-income groups [22]. This may be due to the composition of their sample (which was meant to over-represent extreme ends of the socioeconomic spectrum), changes in typical U.S. parenting practices since the 1980's (which may have become more pedagogical and child-focused across SES groups), or the way their input measures were conceptualized and sampled (only CDS from the primary caregiver). Any of these factors might be responsible for over (or under) estimation of typical speech exposure. A recent word-learning model, incorporating additional data from a wider range of socioeconomic groups, as well as different speakers and types of speech (i.e., child-directed vs. overheard) produced a more conservative estimate of roughly 1200 word tokens/hour in CDS and 1800 word tokens/hour in overheard speech [23]. Assuming a shorter 12-hour speech day, they estimated children would hear about 5.3 million child-directed tokens per year (or 442 speech hours/year per our calculation) and 7.9 million overheard tokens per year (or 658 speech hours/year).

Secondly, the sample Hart & Risley drew from represents a very small percentage of households. Studies of input have found that speech exposure varies wildly across populations. To address this problem, we can use estimates from meta-analyses and systematic reviews, which give us tools to average across studies and identify outlying data. The previously cited meta-analysis examining socioeconomic differences in input found that mean speech exposure based on 10 studies ranged from 1429 to 1956 tokens per hour (522 to 714 speech hours/year assuming 12-hour days) across socioeconomic groups.

Another recent systematic review examined estimates of child-directed vocalization to see how much this varied across human populations in other studies in which children were observed over long periods of time [24]. In that dataset, the population in which child-directed vocalizations were least prevalent was the Pirahã, an indigenous hunter-gatherer tribe in Brazil[5]. The review estimated frequency of child-directed vocalizations at only .5% of total observations.

Many of the previous estimates could still be affected by the presence of an experimenter. For instance, parents might

[4] There may also be some cultural differences, with one paper finding two hours of difference in night-time sleep across two countries [19]. Since these are smaller than individual variation, we do not integrate them into our estimates below.

[5] P. Gordon, Z. Li, S. Medeiros, J. Tang, E. Kirbyn, and D. Everett, "Optimal learning from minimal input: How Pirahã infants acquire language." Presented at Many Paths to Language, 2020.

Table 1: *Estimates of the proportion of awake time and number of input hours accumulated over a year (assuming infants are awake 8 or 15 h / day). "Optimal" refers to quality-wise ideal input that is child-directed, by mother, without background noise, and no speaker overlap. Each row indicates how much input gets added as one of the "optimal" conditions is no longer met; e.g., the "Non-directed" % min indicates that in this condition, approximately 3% of the infants' awake time is speech that is not-directed to the child. Two estimates are shown for each condition: minimum, where the most conservative assumptions in Section 2 with the minimum experience recorded in Section 3 are applied, and maximum with the most liberal assumptions and maximal experience.*

Qualities	**Proportion of waking hours**		**Annual input with assumed waking hours**			
	% min	% max	8h min	8h max	15h min	15h max
Optimal	0.04	26.46	1.17	773.16	2.19	1449.68
Overlapping	0.005	2.94	0.15	85.91	0.27	161.08
Noisy	0.105	0.00	3.07	0.00	5.75	0.00
Non-mother	0.35	12.60	10.23	368.17	19.18	690.32
Non-directed	3.071	18.00	89.73	525.96	168.25	986.18
Total	3.571	60.00	104.35	1753.2	195.64	3287.26

talk more to their child because they think it is desirable to the researchers. Alternatively, they may speak less or in a different style (e.g., different speaking rate or vocabulary) because they are uncomfortable with the situation (see Exp. 2 in [7]).

Fortunately, modern technology enables less obtrusive recordings of everyday language experiences of children: long-form recordings from microphones worn by children throughout a typical day [25]. These recordings are usually between 8 to 16h in duration, continuously capturing everything the child hears. Since this recording takes place without the presence of an experimenter, more ecologically plausible estimates of input can be derived from long-form audio.

A recent cross-linguistic study by Bunce et al.[2] analyzed carefully hand-annotated sections of long-form audio in North-American and UK English, Argentinian Spanish, Tseltal, and Yélî Dnye. They separately analyzed adult speech addressed at the target child, at other adults, or at other children. The maximum total input, summing across all addressee types, was for Yélî Dnye, averaging at 35.7 min of speech per hour heard by the child. We note that this estimate actually counts overlapping signal as additional (i.e., if three people's vocalizations overlap for 1 second, those researchers counted 3 seconds of speech). Even if it is arguable that such heavily overlapping input is as useful as non-overlapping, we include this estimate for our upper-bound, with the least restrictive assumptions.

One challenge of using long-form audio is that it is difficult (if not impossible) to infer when the child is awake even if speech is present in the vicinity of the child. Another challenge is that the current recordings do not usually cover a full 24-hour day, but tend to be biased towards daytime. Thereby it is unclear how to extrapolate the observed patterns of parental speech behavior to full days. Parents of young children know well that only some children sleep throughout that time, and that there are some potential times for exposure in between those hours. In our present estimates, we will simply assume that the rate of speech input is stable across all the time the child is awake, even if it takes place in the middle of the night.

4. Estimation

To estimate the minimum bound of experience, we take all of our most conservative estimates discussed in Section 2, combined with the minimal quantitative information summarised in Section 3. The minimum quantity of child-directed vocalizations corresponds to the Pirahã, for whom all and only child-directed vocalizations were counted. We then subset to 30% of speech as coming from the mother, 30% noise-free, 10% containing overlap. For the maximum bound, we estimate the proportions of each input type using the maximum observed quantities from Yélî Dnye's speech data and the most liberal assumptions outlined in Section 2. Then we multiply both of those sets of proportions by the minimum number of awake hours found in previous work, which is 8 hours awake per day, and the maximum, 15h, times 365 days a year. Results are shown in Table 1.

This estimation work suggests that based on observational data and reasonable assumptions, as little as .04% of infants' awake time may include so-called optimal speech input: speech that is directed to the child, does not overlap with noise or other voices, and is spoken by the mother. In the best plausible case, whereby infants can use all input (the so-called optimal, but also from overlapping voices, speech overlaid with noise, both the mother and others, and non-child-directed speech), then about 3.5% of infants' awake time would contain speech input. Based on the most liberal estimates, optimal speech input would be available for 26% of infants' awake time, and all input would constitute 60% of awake time. These leads to a most conservative estimate of 1 or 2 hours of optimal input in a year, with the most liberal estimate being up to 3,300 hours of total input in the same time.

5. Issues and limitations of our estimates

5.1. Limited to measures of input quantity

The estimates provided above primarily deal with measures of input quantity, or how much speech children are exposed to. In addition, certain aspects of input quality were also considered indirectly, via child-directed speech and the presence of background noise. But the quality of input goes far beyond what we have considered here, ranging from linguistic properties like grammatical complexity and the diversity of words used, to behavioral properties like responsiveness to child speech and integration with non-verbal cues such as gesture and attention [26]. The quality of speech may influence its learnability, which would impact our estimates of how much "effective" input children are receiving. While these considerations are beyond the scope of the current paper, future work should consider these measures further.

5.2. Limited data on learning in non-Western countries

A continuing challenge of estimating infant speech exposure is the relatively little work that has been done outside of the U.S. and European countries [5]. Some of our reasoning is based

on theories that assume what is typical in these countries, but there are many reasons to believe that cultural differences in the linguistic environments of children growing up in rural non-Western countries may affect what kinds of input enter into the learning process. For example, the assumption that children might prefer maternal speech stems from research conducted primarily in households with small nuclear families, where the mother is typically the primary caregiver and by extension children's primary source of speech. But in other cultures, multiple alloparental caregivers provide as much as 50% of total early childcare [27]. Children also spend more of their time with older siblings, who also contribute to caregiving activities, resulting in more child-directed speech coming from other children than from adults [7, 28]. These differences may affect what a child might take in as meaningful input, and as a result change our estimations of speech exposure.

5.3. Unable to incorporate cumulative developments in child language

An additional limitation of these estimates is that modeling children's learning requires knowledge of how learning changes across the early lifespan. Some of these changes result directly from children's growing linguistic knowledge. As children's vocabulary and grammatical knowledge grows, they are more easily able to learn from complex or abstract speech [29]. There is also evidence that lexical processing is itself affected by prior input, such that young children with greater early speech exposure come to recognize words more quickly, and thus can learn more from speech later on [30]. Finally, internationally-adopted preschoolers with no prior English exposure learn words at roughly four times the rate of vocabulary-matched non-adopted infants, suggesting either age or prior L1 (Chinese/Russian) exposure increases ability to learn from speech [31]. These findings all suggest that children's effective input changes as children mature and understand more of their language.

6. Discussion

The amount of infant speech exposure has consequences for models that try to explain the human learning process. A plausible model of learning should be able to achieve infant-like language comprehension skills from input that is also comparable to infant input in quantity (and ideally in quality). The limits identified above are an attempt to provide a scale of input for which age-appropriate learning should be possible, but excluding other factors in learning, such as interaction, feedback, and attunement of input to the child's developmental level.

Despite our efforts, many uncertainties remain in these estimates, some of which we have attempted to make transparent in this paper. This means that the amount of speech input that truly counts for infant learning remains unclear. What can be quite confidently stated is that modeling studies substantially exceeding the upper limit (approx. 3300 h or $\sim$40M word tokens per year) should be interpreted with caution, as little empirical support for such input quantities exists.

Beyond the uncertainties involved in getting the average exposure quantity right, there is known to be substantial individual variability in language input between different families and communities, with previous estimates relying on averages across many children. At the same time, substantial individual variability exists in early learning outcomes, as exemplified by differences in receptive and productive vocabulary sizes during the first years of life (e.g., Child Development Inventories data in Wordbank; [32]). Thereby, the assumption is not that all reasonable amounts of input data should result in comparable learning outcomes, be it for models or real infants. However, the exact relation between input and outcomes is not clear. Hart & Risley found that caregiver speech was highly correlated with child vocabulary (r = .73), such that input quantity explained approximately half of total variance in languages outcomes, and suggested a fairly linear relationship between input and outcome (i.e., children who hear twice as much speech also produce twice as many different words in naturalistic settings)[20]. A recent meta-analysis found no evidence that input-outcome associations differ in size in higher- vs. lower-input samples[6], further suggesting a linear relationship, but also a much smaller estimated effect of input (r = .23, or only 5% variance).

Still, other studies report different associations between input and outcome. One study [8] also found a large (r = 0.57) correlation between amount of CDS input and vocabulary size. However, the differences in input were on the order of magnitude, whereas the explained differences in vocabulary sizes differed approximately by a factor of two only. Additional evidence comes from children growing up in rural, non-Western settings, who reportedly experience anywhere from $1/4$ to $1/10$ the amount of input as children in urban, mostly Western settings [24]. While studies have found that many children in these settings may be at increased risk of disruption in cognitive development [33], there is no evidence that these children systematically encounter language delays of this magnitude. In fact, several studies have found that children in these settings attain key language milestones at roughly the same time as children growing up in the U.S. [34, 35].

For modeling experiments, all this suggests that models should exhibit learning differences for different amounts of inputs. However, they should not exhibit vocabulary growth differences of several orders of magnitude even if the input does so. On the other hand, typical machine learning algorithms—the basis for contemporary learning models—already have diminishing returns with increasing data, where performance improves approximately linearly with respect to logarithm of the data amount. Yet, it is possible that there are also various non-linearities involved with both humans and computational models, such as the need to amass a certain quantity of data before being able to make certain morphosyntactic generalizations, or to reach qualitatively leaps in representing the elements and compositional structure of spoken language. Alternatively, learning outcomes that appear as non-linear with respect to language experience may not always be such (see, e.g., [36]).

To conclude, the possible range of typical speech input quantity is broad, and more research is needed to reduce the uncertainties involved in these estimates. In parallel, more empirical research is needed to understand how variability in input should map onto learning outcomes. Both accurate input and outcome measures are needed to create more plausible models of early language acquisition.

7. Acknowledgements

O.R. was funded by L-SCALE project from Kone Foundation, and A.C. by the J.S. McDonnell Foundation Understanding Human Cognition Scholar Award; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant No. 101001095).

[6]Coffey & Snedeker, preprint available from `https://osf.io/aydcf/`

8. References

[1] T. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[2] R. Algayres, Y. Adi, T. Nguyen, J. Copet, G. Synnaeve, B. Sagot, and E. Dupoux, "Generative spoken language model based on continuous word-sized audio tokens," in *Proc. EMNLP-2023*, Singapore, Dec. 2023, pp. 3008–3028.

[3] M. Lavechin, Y. Sy, H. Titeux, M. Blandon, O. Räsänen, H. Bredin, E. Dupoux, and A. Cristia, "Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models," in *Proc. INTERSPEECH 2023*, Incheon, Korea, Sep. 2023, pp. 4588–4592.

[4] A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, "Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora," in *Proc. CoNLL-2023*, Singapore, Dec. 2023, pp. 1–34.

[5] E. Kidd and R. Garcia, "How diverse is child language acquisition research?" *First Language*, vol. 42, no. 6, pp. 703–735, 2022.

[6] N. F. Ramírez, S. Lytle, and P. Kuhl, "Parent coaching increases conversational turns and advances infant language development," *Proceedings of the National Academy of Sciences*, vol. 117, no. 7, pp. 3484–3491, 2020.

[7] L. A. Shneidman and S. Goldin-Meadow, "Language input and acquisition in a mayan village: How important is directed speech?" *Developmental Science*, vol. 15, no. 5, pp. 659–673, 2012.

[8] A. Weisleder and A. Fernald, "Talking to children matters: Early language experience strengthens processing and builds vocabulary," *Psychological Science*, vol. 24, no. 11, pp. 2143–2152, 2013.

[9] D. Sperry, L. Sperry, and P. Miller, "Reexamining the verbal environments of children from different socioeconomic backgrounds," *Child Development*, vol. 90, no. 4, pp. 1303–1318, 2019.

[10] N. Pancsofar and L. Vernon-Feagans, "Mother and father language input to young children: Contributions to later language development," *Journal of Applied Developmental Psychology*, vol. 27, no. 6, pp. 571–587, 2006.

[11] L. Vernon-Feagans and M.E. Bratsch-Hines and Family Life Project Key Investigators and others, "Caregiver–child verbal interactions in child care: A buffer against poor language outcomes when maternal language input is less," *Early Childhood Research Quarterly*, vol. 28, no. 4, pp. 858–873, 2013.

[12] T. S. Duncan and J. Paradis, "Home language environment and children's second language acquisition: The special status of input from older siblings," *Journal of Child Language*, vol. 47, no. 5, pp. 982–1005, 2020.

[13] A. Neuman, M. Wroblewski, J. Hajicek, and A. Rubenstein, "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear and Hearing*, vol. 31, no. 3, pp. 336–344, 2010.

[14] K. G. Riley and K. McGregor, "Noise hampers children's expressive word learning," *Language, Speech, and Hearing Services in Schools*, vol. 43, no. 3, pp. 325–337, 2012.

[15] S. Stansfeld, B. Berglund, C. Clark, I. Lopez-Barrio, P. Fischer, E. Öhrström, M. Haines, J. Head, S. Hygge, I. van Kamp, and B. Berry, "Aircraft and road traffic noise and children's cognition and health: a cross-national study," *The Lancet*, vol. 365, no. 9475, pp. 1942–1949, 2005.

[16] M. Lavechin, M. Métais, H. Titeux, A. Boissonnet, J. Copet, M. Rivière, E. Bergelson, A. Cristia, E. Dupoux, and H. Bredin, "Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and c50 room acoustics estimation," in *Proc. IEEE ASRU-2023*, Taipei, Taiwan, Dec. 2023.

[17] O. Räsänen et al., "Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech," *Speech Communication*, vol. 113, pp. 63–80, 2019.

[18] B. C. Galland, B. J. Taylor, D. E. Elder, and P. Herbison, "Normal sleep patterns in infants and children: a systematic review of observational studies," *Sleep Medicine Reviews*, vol. 16, no. 3, pp. 213–222, 2012.

[19] J. A. Mindell, A. Sadeh, B. Wiegand, T. H. How, and D. Y. Goh, "Cross-cultural differences in infant and toddler sleep," *Sleep Medicine*, vol. 11, no. 3, pp. 274–280, 2010.

[20] B. Hart and T. Risley, *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing, 1995.

[21] M. Brent and J. Siskind, "The role of exposure to isolated word in early vocabulary development," *Cognition*, vol. 81, no. 2, pp. 31–44, 2001.

[22] S. Dailey and E. Bergelson, "Language input to infants of different socioeconomic statuses: A quantitative meta-analysis," *Developmental Science*, vol. 25, no. 3, p. e13192, 2022.

[23] G. Kachergis, V. Marchman, and M. Frank, "Toward a "standard model" of early language learning," *Current Directions in Psychological Science*, vol. 31, no. 1, pp. 20–27, 2022.

[24] A. Cristia, "A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations," *Developmental Science*, vol. 26, p. e13265, 2023.

[25] M. Lavechin, M. De Seyssel, L. Gautheron, E. Dupoux, and A. Cristia, "Reverse engineering language acquisition with child-centered long-form recordings," *Annual Review of Linguistics*, vol. 8, pp. 389–407, 2022.

[26] M. Rowe and C. Snow, "Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual," *Journal of Child Language*, vol. 47, no. 1, pp. 5–21, 2020.

[27] N. Chaudhary, G. Salali, and A. Swanepoel, "Sensitive responsiveness and multiple caregiving networks among mbendjele bayaka hunter-gatherers: Potential implications for psychological development and well-being," *Developmental Psychology*, vol. 60, no. 3, pp. 422–440, 2023.

[28] G. Loukatou, C. Scaff, K. Demuth, A. Cristia, and N. Havron, "Child-directed and overheard input from different speakers in two distinct cultures," *Journal of Child Language*, vol. 49, no. 6, pp. 1173–1192, 2022.

[29] M. L. Rowe, "A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development," *Child Development*, vol. 83, no. 5, pp. 1762–1774, 2012.

[30] N. Hurtado, V. Marchman, and A. Fernald, "Does input influence uptake? links between maternal talk, processing speed and vocabulary size in spanish-learning children," *Developmental Science*, vol. 11, no. 6, pp. F31–F39, 2008.

[31] J. Snedeker, J. Geren, and C. Shafo, "Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of english in internationally-adopted children," *Cognitive Psychology*, vol. 65, no. 1, pp. 39–76, 2012.

[32] M. C. Frank, M. Braginsky, D. Yurovsky, and V. A. Marchman, "Wordbank: An open repository for developmental vocabulary data," *Journal of Child Language*, vol. 44, no. 3, pp. 677–694, 2017.

[33] A. Sania, C. Sudfeld, and G. Danaei, "Early life risk factors of motor, cognitive and language development: a pooled analysis of studies from low/middle-income countries," *BMJ Open*, vol. 9, p. e026449, 2019.

[34] M. Casillas, P. Brown, and S. Levinson, "Early language experience in a tseltal mayan village," *Child Development*, vol. 91, no. 5, pp. 1819–1835, 2019.

[35] M. Casillas, P. Brown, and S. C. Levinson, "Early language experience in a papuan community," *Journal of Child Language*, vol. 48, no. 4, pp. 792–814, 2021.

[36] B. McMurray, "Defusing the childhood vocabulary explosion," *Science*, vol. 317, no. 5838, pp. 631–631, 2007.