What do Babies hear? Analyses of Child- and Adult-Directed Speech

Marisa Casillas[1], *Andrei Amatuni*[2], *Amanda Seidl*[3], *Melanie Soderstrom*[4], *Anne S. Warlaumont*[5], *Elika Bergelson*[2]

[1]Max Planck Institute for Psycholinguistics, The Netherlands
[2]Psychology and Neuroscience, Duke University, USA
[3]Speech, Language, and Hearing Sciences, Purdue University, USA
[4]Psychology, University of Manitoba, Canada
[5]Cognitive and Information Sciences, University of California, Merced, USA

marisa.casillas@mpi.nl, andrei.amatuni@duke.edu, aseidl@purdue.edu, m_soderstrom@umanitoba.ca, awarlaumont2@ucmerced.edu, elika.bergelson@duke.edu

Abstract

Child-directed speech is argued to facilitate language development, and is found cross-linguistically and cross-culturally to varying degrees. However, previous research has generally focused on short samples of child-caregiver interaction, often in the lab or with experimenters present. We test the generalizability of this phenomenon with an initial descriptive analysis of the speech heard by young children in a large, unique collection of naturalistic, daylong home recordings. Trained annotators coded automatically-detected adult speech 'utterances' from 61 homes across 4 North American cities, gathered from children (age 2-24 months) wearing audio recorders during a typical day. Coders marked the speaker gender (male/female) and intended addressee (child/adult), yielding 10,886 addressee and gender tags from 2,523 minutes of audio (cf. HB-CHAAC Interspeech ComParE challenge; Schuller et al., in press). Automated speaker-diarization (LENA) incorrectly gender-tagged 30% of male adult utterances, compared to manually-coded consensus. Furthermore, we find effects of SES and gender on child-directed and overall speech, increasing child-directed speech with child age, and interactions of speaker gender, child gender, and child age: female caretakers increased their child-directed speech more with age than male caretakers did, but only for male infants. Implications for language acquisition and existing classification algorithms are discussed.

Index Terms: Addressee, Child Directed Speech, Language Development, Speech Classification, Gender

1. Introduction

Speech directed to infants and young children often has particular linguistic and acoustic characteristics that differ from those of adult-directed speech [1]. These characteristics, and the caregiver-infant interactions that accompany them, are hypothesized to play a critical role in language development [2]. Indeed, recent findings suggest that greater exposure to child-directed speech (CDS), but not adult-directed speech (ADS), is related to faster lexical processing and larger vocabularies in toddlers [3]. Speech directed to children from adults appears to have an impact on early lexical development even in cultures where adults address children infrequently [4]. To date, the vast majority of the research on CDS has relied on relatively short, constrained recordings, predominantly of mother-infant interactions, which are unlikely to capture the kinds of speech that infants hear in their daily lives (cf. [3, 4, 5]. Emerging technologies like LENA [6], which allow for automated analysis of full-day real-world recordings of infants' language experiences, open the door for more ecologically valid analysis. But, to date, automated approaches to classifying adult- and child-directed speech have been limited [7, 8, 9], in large part due to insufficient quantities of tagged, appropriate, and diverse data on which classifiers can be trained.

In the current study we analyze adult speech from a collection of "daylong" recordings of young children, their caregivers, and other family members in their natural home environment. The data were annotated with an eye towards developing CDS and ADS classifiers. Such algorithms, in turn, can be used over thousands of hours of existing data, with great potential for improving our understanding of infant learning and parent-child interaction, both for typically-developing children (such as those in the present dataset), and by extension, cross-culturally and to special populations. We report here on dataset construction and initial analysis of the manually annotated data.

Our main questions of interest were: (1) the accuracy of LENA's automated tags, (2) the role of socioeconomic status (SES) and gender on CDS and ADS, and (3) changes in CDS and ADS over the first two years of life.

2. Methods

This dataset was created by sub-sampling daylong audio recordings from four corpora that come from a larger repository of real-world child language recordings, HomeBank ([10] homebank.talkbank.org): Bergelson [11], McDivitt [12], VanDam [13], and Warlaumont [14]. All recordings were collected from typically-developing children with a LENA audio recorder, which was worn by the child in specialized clothing [6]. The recordings were made with families from four North American cities who primarily use English at home. All families granted permission to share the audio with the research community. We selected one recording from each of 61 children who met the description above, sampling as uniformly as possible between 0 and 2 years across the combined corpora (Figure 1). Though the corpora came from four labs with different research questions driving their data collection, a unified annotation system allowed for interoperability, pooled coding, and analysis.[1]

The recordings were first analyzed with LENA's proprietary software, which identifies "conversational blocks' (i.e. speech

[1] Each corpus had a different profile for the age and socioeconomic status of its participants but, notably, the McDivitt corpus focused specifically on young mothers.

surrounded by 5 s of non-speech) and utterance boundaries, adding speaker tags from a closed set of 12 alternatives (e.g. Female-Adult-Near, Target-Child) and some other features. Our present goal was to subsample evenly over early childhood, and classify the adult speech in these recordings as (1) spoken by a female or male and (2) addressed to a child or adult listener.

We selected 20 of LENA's conversational blocks from each of the 61 recordings, only including blocks with at least 10 Female-Adult-Near (FAN) or Male-Adult-Near (MAN) speaker tags, as determined by the LENA system. The 20 blocks for each recording were then extracted from the audio file and spliced into their component utterances (created by LENA's proprietary diarization algorithm; hereafter, known as 'clips'). Custom client and server software[2] was written to randomly distribute the blocks, have annotators listen to and label the individual clips, and then send back their labeled responses to the server, which would remove that block from the remaining pool for that particular annotator. Each block was distributed to three different trained annotators for independent labeling; all nine annotators first completed a training data set not included in the present analysis.

Annotators' gender and addressee tags relied on both acoustic-phonetic information and context (see `https://osf.io/d9ac4/` for details). Non-speech and ambiguous speech clips were tagged into as "Junk". Annotators achieved high reliability in differentiating CDS/ADS (Fleiss' kappa $>$ 0.75, $p < 0.001$). This same set of clips and tags is being used for the ComParE 2017 addressee sub-challenge (HomeBank Child/Adult Addressee Sub-Challenge, HB-CHAAC).

3. Results

Three annotators tagged each of the 12,684 MAN and FAN speech clips detected across the 3,660 conversational blocks by LENA's proprietary software. Each clip was then assigned a 'true' value for speaker gender and addressee if at least two of the three annotators converged on a single decision (gender: Male, Female, or Junk; addressee: CDS, ADS, Junk). In the case that annotators' judgments were split evenly, the clip was labeled as 'no-majority'. The findings reported below first briefly assess LENA's accuracy in detecting speech and assigning gender labels and then turn to the analysis of human-coded adult- and child-directed speech with respect to child age, child and adult gender, and SES.

3.1. LENA label accuracy

Using our manual annotations with majority consensus across coders as the gold standard, we first assessed false positives from the LENA gender tags (FAN and MAN). 1,730 (14%) of the clips LENA tagged as FAN or MAN were classified as "Junk", meaning that the clip either did not contain speech, had no identifiable primary speaker, or was too ambiguous to make a decision. Thus, reliance on LENA's labels alone may lead researchers to overestimate the quantity of speech in the child's environment. That said, the nature of this dataset lets us find false positives, but not false alarms; we do not know how many clips *should have* been classified as FAN or MAN but were not. Estimates from a similar set of recordings [3] found a similar false alarm rate of 18%. False alarms should thus be kept in mind when interpreting LENA's label accuracy for speech clips (cf. [15]).

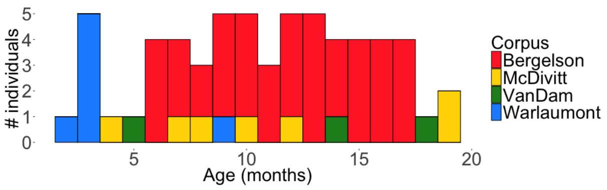

Figure 1: *Age and corpus distribution for the 61 children's recordings included in the HB-CHAAC sample.*

Comparing the accuracy of LENA's gender labels with those of our trained annotators, LENA's MAN clips turned out to be male 67% of the time and female 26% of the time. Even ignoring Junk and non-majority clips, MAN clips only turn out to be male speech 72% of the time (the rest being female speech). This likely relates to the fact that female speech is far more frequent in the infants' environment in our dataset and others' [16]; here it outstrips male speech nearly three to one. As others have found [15], LENA speaker-tag errors are systematic: women were more likely to be tagged as men when they used ADS (MAN tags for female CDS: 7%; for female ADS: 16%) and men were more likely to be tagged as women when they used CDS (FAN tags for male CDS: 29%; for male ADS: 9%). So again, based on LENA's labels alone, researchers might overestimate the quantity of male speech in the child's environment, though the false alarm rate here is unknown.

3.2. Child- and adult-directed speech

Putting LENA's automated labels aside, we next analyzed our human-verified annotations of the speaker's gender and addressee. The dataset included 10,861 speech clips with a majority human code for both gender and addressee, 2,776 of which came from male adult speakers, $\sim$60% of which were CDS (for male and female speakers alike). We first investigated proportional CDS patterns—how much of the speech children hear is directed toward them?—and then we looked at overall speech quantity—how much speech do children hear overall?

3.2.1. *Proportion of CDS by gender*

We first looked at how CDS rates changed with children's age, splitting our analysis on the gender of the child and the gender of the adult speaker (Figure 2), given previous research suggesting gender contributes to CDS patterns [16].

We modeled the likelihood that a clip was child-directed (1 or 0) with a mixed-effects logistic regression including fixed effects of child age (in months; centered), child gender (male/female), adult gender (male/female), and full interactions between these three predictors, along with nested random effects of corpus (Bergelson/McDivitt/VanDam/Warlaumont) and child (child ID).[3] We found that likelihood of CDS increased with age ($\beta = 0.15$, $SE = 0.05$, $z = 3.02$, $p = 0.0025$), that women were overall more likely to use CDS compared to men ($\beta = 0.32$, $SE = 0.09$, $z = 3.61$, $p = 0.0003$), and that women's CDS to boy infants increased significantly more with age compared to men's ($\beta = 0.14$, $SE = 0.02$, $z = 5.42$, $p < 0.0001$).

We next constructed a parallel model on these same data (i.e., human-verified instances of adult speech), this time using LENA's MAN and FAN labels in lieu of our human-verified gender labels. The LENA-based model results were qualita-

[2] https://github.com/SeedlingsBabylab/idslabel

[3] Including child is equivalent to adding a random effect for recording because there is one recording per child.

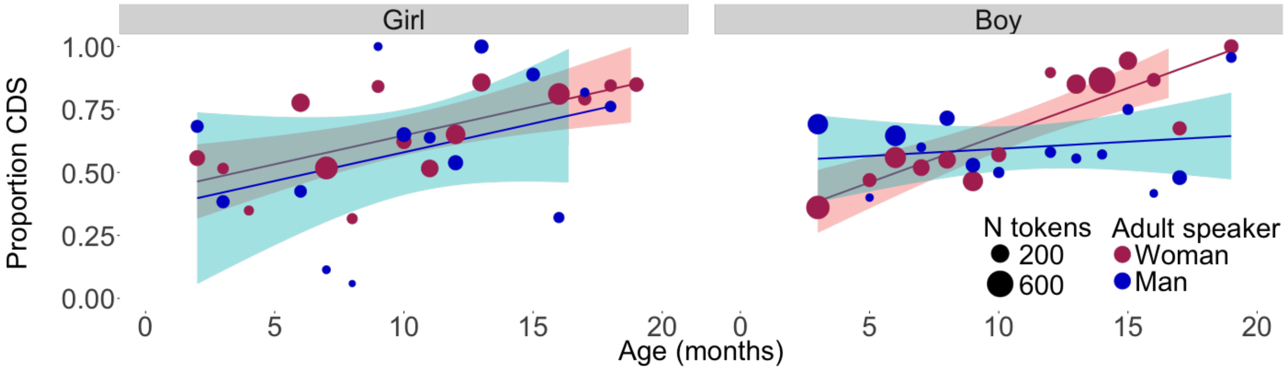

Figure 2: *Proportion of adult CDS across age, grouped by child and adult gender.*

tively similar, showing an increase in CDS with age (β = 0.15, *SE* = 0.05, z = 2.96, p = 0.003), more CDS for FAN speakers than MAN speakers (β = 0.75, *SE* = 0.08, z = 9.34, $p < 0.0001$), and the same three-way interaction showing greater gains with age for CDS from FAN to boy infants (β = 0.05, *SE* = 0.02, z = 2.06, p = 0.039). Thus our results for use of CDS are similar to those with LENA's automated labels, but the LENA-based model underestimates the size of the three-way interaction and overestimates the overall effect of more CDS from women than men.

3.2.2. *Proportion of CDS by SES*

We next examined SES, using maternal education as our proxy; this was the only individual index of SES we had for all children across the four corpora, and has been linked to language development in previous research [17, 18]. We then evaluated CDS with respect to whether the child's mother had attained a university-level degree or not (no degree = 18 (29%) of children; Figure 3).

We again used mixed-effects logistic regression to model the likelihood that a clip was child-directed (1 or 0) with fixed effects of child age (in months; centered), maternal education (university degree/not), and their interaction, plus nested random effects of corpus (Bergelson/McDivitt/VanDam/Warlaumont) and child (child ID). As before, we found a robust increase in CDS with age (β = 0.20, *SE* = 0.07, z = 2.89, p = 0.004), but no effect of education (β = 0.22, *SE* = 0.38, z = 0.57, p = 0.56) and no interaction of child age and maternal education (β = -0.06, *SE* = 0.08, z = -0.75, p = 0.45). In short, we found little evidence that maternal education affects the likelihood that any single clip is child-directed, bearing in mind that our binary measure and limited variance in SES restrict our insight.

3.2.3. *Quantity of speech by gender*

We next checked for changes in the sheer quantity of speech directed toward children. Because our dataset samples exactly 20 conversational blocks from each child, we can directly compare differences in the overall amount of speech individual children heard during their recordings.

We first analyzed the quantity of speech children heard, given child age, child gender, and adult gender. Our dependent variable was then the total number of human-validated speech tokens in our dataset heard by each child from either a male or female adult speaker. This set of analyses then aggregates our dataset down to 114 datapoints.[4] We modeled the number of speech clips children heard with a mixed-effects linear regression including fixed effects of child age (in months; centered), child gender (male/female), adult gender (male/female) and full interactions between these three predictors, plus a random effect of corpus (Bergelson/McDivitt/VanDam/Warlaumont) and child (child ID). Female speakers contributed significantly more speech tokens than males (β = 66.88, *SE* = 21.11, t = 3.17) but there were no other significant effects on raw number of speech children heard, including no effects of child age, child gender, or their interactions with adult gender (all $|t| < 1.75$).

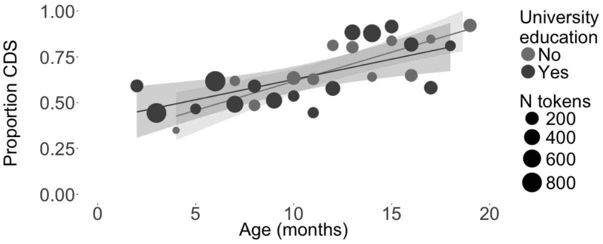

Figure 3: *Proportion of CDS across age, grouped by maternal education.*

Zooming in to just CDS, we modeled the total number of CDS tokens children heard using another regression with the same fixed- and random-effects structure.[5] As before, the only statistically significant effect was that female speakers contributed significantly more speech tokens than male ones (β = 46.65, *SE* = 14.59, t = 3.20; all other predictors had $|t| < 0.9$). Thus, predictors for the proportion of CDS heard by children (age and parent/child gender) were not predictive of overall speech quantity or overall CDS quantity.

3.2.4. *Quantity of speech by SES*

Turning to SES, we next assessed the quantity of speech children heard, split on child age and adult education (Figure 4). Our dependent variable was the total number of human-validated speech tokens in our dataset heard by each child, aggregating our dataset to 61 datapoints. We modeled the number of speech clips children heard with an-

[4] Up to two datapoints for each child (number of clips from men and women). Note that eight children heard no male speech in our dataset.)

[5] We excluded one outlying 14-month-old boy whose number of CDS tokens was greater than 3 *SD*s from the mean.

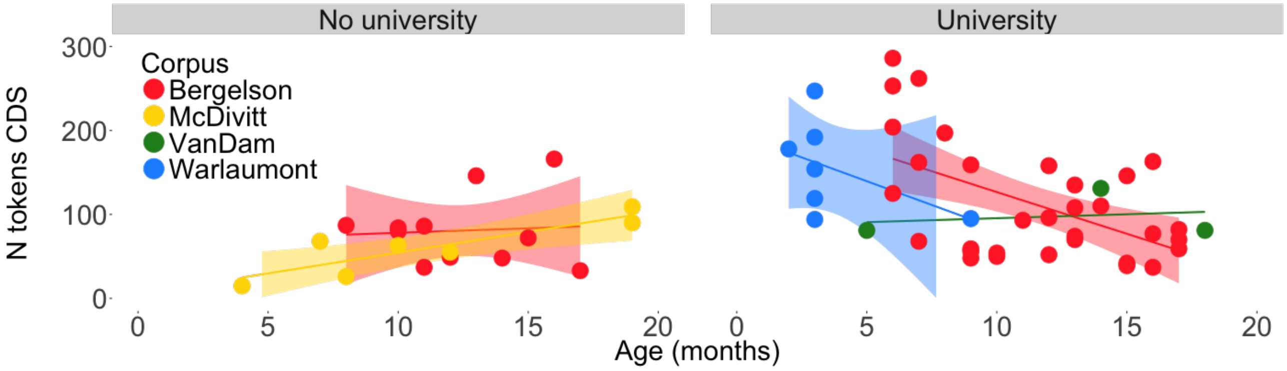

Figure 4: *Total number of CDS tokens found within the 20 analyzed conversation blocks for each child, plotted here with respect to child age, whether infants' mothers attained a university degree, and which corpus the data came from (one outlier removed).*

other linear regression with fixed effects of child age (in months; centered), maternal education (university degree/not), and their interaction, plus a random effect of corpus (Bergelson/McDivitt/VanDam/Warlaumont).[6] There was no effect of child age (β = 2.02, *SE* = 4.79, t = 0.42), but recordings in households with university-educated mothers showed significantly more speech than houses without (β = 94.37, *SE* = 23.58, t = 4.00), with a significant interaction between child age and maternal education (β = -20.44, *SE* = 5.49, t = -3.72). Intriguingly, we find that children with university-educated mothers experienced a decrease in total amount of speech with age, while children with non-university-educated mothers experienced a stable amount of speech. We return to this in the discussion.

To see whether this pattern applied to the raw number of CDS clips, we ran a final regression using the same fixed- and random-effects structure, with the same excluded outlier as before. Again there was no overall effect of child age (β = 4.13, *SE* = 3.10, t = 1.33), but significant effects of maternal education and an interaction between child age and maternal education, mirroring the previous model (β = 46.54, *SE* = 15.35, t = 3.03 and β = -11.35, *SE* = 3.56, t = -3.19, respectively). Therefore the decrease in speech tokens with age for university-educated mothers also holds for number of CDS tokens alone.

4. Discussion and Conclusions

In sum, we find that over early development, parents increase the proportion of CDS in infants' input, with male children hearing more CDS from female caregivers than male ones as they get older. We also find that maternal education does *not* predict variability in the *proportion* of CDS over time, but *does* lead to a global difference in *quantity* of speech and specifically child-directed speech, though this effect decreases over time.

Our models using LENA's gender tags were qualitatively simiar to human tags (though with different effect sizes), despite relatively high error-rates in LENA's MAN tags. We did not have automated CDS/ADS tags (cf. 2017 HB-CHAAC ComParE Challenge), but hope that advances in machine learning will provide increasingly aligned human-machine addressee and gender annotation, which could then be confidently applied to the growing mass of daylong recordings (e.g. in Homebank). Iteratively, this will increase the quantity of analyzable data, thereby expanding the generalizability of the present results.

While we find convergent average rates of CDS with previous work with U.S. and non-Western infants [4], our finding that CDS proportion increases over this age range is a novel contribution [19]. Our child- and speaker-gender results, which showed that female speakers are more responsive to male infants, diverge somewhat from previous work [16], which found that female speakers respond more to female infants than male infants. Given that that work used only LENA tags, and infants <8 months, further research is necessary to explore these differences.

Our results also contribute to an increasingly complex understanding of the interaction between SES and language development. We found that the proportion of CDS in the input did not vary with SES, but that infants with less-educated mothers heard less speech overall, *and* less child-directed speech in particular. Notably, the education difference decreased with age, with infants across the dataset hearing similar quantities of CDS and ADS in the second year. This is broadly in-line with previous work [3], though we use somewhat different operationalizations of 'overheard' and 'child-directed' speech.

Taken together, our results provide first steps towards understanding the distribution of CDS in naturalistic conditions, across SES and age within North America. Future work is needed to assess the generalizability of these results globally, hopefully with the help of improved computational tools. This work also provides an example for how shared datasets with interoperable classification schema provide more robust analysis than any lab's data alone.

In sum, what do babies hear? They hear two-thirds of their input in a speech register that is increasingly hand-tailored to them as they become active participants in caregiver interactions over the first two years of life.

5. Acknowledgements

This work was funded in part by NWO Veni Innovational Research Scheme 275-89-033 (MC), NIH DP5-OD019812 (EB), SSHRC Insight Grant 435-2015-0628 (MS), and NSF SBE-1539129 and NSF BCS-1529127 (AW). The authors further thank the research assistants who provided HB-CHAAC labels and Kelsey Dyck for assistance developing the labeling protocol, Mark VanDam for sharing his corpus, Brian MacWhinney hosting the HomeBank corpora, as well as all families who provided the raw data.

[6] With one datapoint per child there was no random effect of child.

6. References

[1] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants," *Developmental Review*, vol. 27, no. 4, pp. 501–532, 2007.

[2] R. M. Golinkoff, D. D. Can, M. Soderstrom, and K. Hirsh-Pasek, "(baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition," *Current Directions in Psychological Science*, vol. 24, no. 5, pp. 339–344, 2015.

[3] 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.

[4] 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.

[5] A. L. Robinson-Mosher and B. Scassellati, "Prosody recognition in male infant-directed speech," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3. IEEE, 2004, pp. 2209–2214.

[6] C. R. Greenwood, K. Thiemann-Bourque, D. Walker, J. Buzhardt, and J. Gilkerson, "Assessing childrens home language environments using automatic speech recognition technology," *Communication Disorders Quarterly*, vol. 32, pp. 83–92, 2011.

[7] S. Schuster, S. Pancoast, M. Ganjoo, M. C. Frank, and D. Jurafsky, "Speaker-independent detection of child-directed speech," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 366–371.

[8] S. Vosoughi and D. K. Roy, "A longitudinal study of prosodic exaggeration in child-directed speech," 2012.

[9] T. Inoue, R. Nakagawa, M. Kondou, T. Koga, and K. Shinohara, "Discrimination between mothers infant-and adult-directed speech using hidden markov models," *Neuroscience research*, vol. 70, no. 1, pp. 62–70, 2011.

[10] M. VanDam, A. S. Warlaumont, E. Bergelson, A. Cristià, P. De Palma, and B. MacWhinney, "Homebank: An online repository of daylong child-centered audio recordings," 2016.

[11] E. Bergelson, "Bergelson Seedlings HomeBank corpus," 2016, doi:10.21415/T5PK6D.

[12] K. McDivitt and M. Soderstrom, "McDivitt HomeBank corpus," 2016, doi: 10.21415/T5KK6G.

[13] M. VanDam, "VanDam2 HomeBank corpus," 2016.

[14] A. S. Warlaumont and G. M. Pretzer, "Warlaumont HomeBank corpus," 2016, doi:10.21415/T54S3C.

[15] M. VanDam and N. H. Silbert, "Fidelity of automatic speech processing for adult and child talker classifications," *PloS one*, vol. 11, no. 8, p. e0160588, 2016.

[16] K. Johnson, M. Caskey, K. Rand, R. Tucker, and B. Vohr, "Gender differences in adult-infant communication in the first months of life," *Pediatrics*, vol. 134, no. 6, pp. e1603–e1610, 2014.

[17] A. Fernald, V. A. Marchman, and A. Weisleder, "Ses differences in language processing skill and vocabulary are evident at 18 months," *Developmental science*, vol. 16, no. 2, pp. 234–248, 2013.

[18] E. Bergelson and D. Swingley, "Young toddlers word comprehension is flexible and efficient," *PloS one*, vol. 8, no. 8, p. e73359, 2013.

[19] J. Huttenlocher, M. Vasilyeva, H. R. Waterfall, J. L. Vevea, and L. V. Hedges, "The varieties of speech to young children." *Developmental psychology*, vol. 43, no. 5, p. 1062, 2007.