

# The LENA Natural Language Study

**Jill Gilkerson & Jeffrey A. Richards**  
LENA Foundation, Boulder, CO

LTR-02-2

September 2008

Software Version: V3.1.0

## ABSTRACT

---

This paper describes the multiphase LENA Natural Language Study, an ongoing data collection effort designed to investigate the language environment of infants and toddlers. Data collected contributes to product development and normative information for use with the LENA System and child development research. Phase I study participants were representative of the US Census with respect to mothers' attained education and consisted of 329 normally developing infants and toddlers from monolingual English-speaking households living in the Denver-metro area. Participants provided day-long audio recordings of their natural language environment once a month, and certified speech language pathologists assessed participant language ability independently through standardized assessments. A subset of 80 Phase I participants has continued to provide monthly recordings in Phase II of the study. The normative database described herein contains over 32,000 hours of spontaneous speech data. This paper describes how normative information was derived for the Adult Word Count estimates (AWC; adult words spoken per day), Conversational Turns estimates (CT; adult-child alternations per day), and Child Vocalization frequency estimates (CV; words, babbles, and "protophones" or pre-speech communicative sounds) that are reported in the LENA System.

### Keyword List

Normative development, language development, language environment, spontaneous speech, Adult Words, AWC, Conversational Turns, CT, Child Vocalizations, CV.

## 1.0 INTRODUCTION

---

### 1.1 Study Purposes

The data collected during the LENA Natural Language Study has resulted in a corpus of spontaneous speech data representative of the language environment of infants and toddlers 2 months – 48 months of age. Daily speech recordings collected during this study provided the basis for the development of the advanced audio processing algorithms central to the LENA System. Following on Hart and Risley's (1995) seminal language development research, these data also were intended to establish normative information about patterns of talk and Adult Word Counts (AWC), Child Vocalizations (CV), and Conversational Turns (CT) in the households of infants and toddlers and to validate the earlier Hart and Risley research.<sup>1</sup>

### 1.2 Study Overview

The LENA Natural Language Study is an ongoing multiphase data collection effort. The current paper describes Phase I, the normative data collection phase which took place between January – June 2006, and the first 18 months of Phase II, an extended longitudinal data collection phase started in July, 2006. At the onset of Phase I we recruited parents of infants and toddlers (predominantly 2 months – 36 months of age, with an additional 15 children 37 to 48 months) through advertisements in local newspapers and direct mail solicitation. Potential participants were selected based on demographic considerations such as the child's age and the mother's education level. Participating families in Phase I provided day-long audio recordings once per month for six months and visited LENA's child language research center for a standard evaluation by a certified speech language pathologist. We compiled speech recording data into the LENA Natural Language Corpus from which we produced normative estimates of daily AWCs, CTs, and CVs in the language environment of infants and toddlers. Phase I audio data have been supplemented by additional audio recordings during Phase II to provide normative information for children up to 48 months old. The LENA Natural Language Study has been reviewed and approved by Essex Institutional Review Board (IRB) to help ensure that the rights and welfare of research participants were protected and that the study was conducted in an ethical manner.

---

1 See <http://www.lenababy.com/Study.aspx> for more information about Hart and Risley (1995) and The Power of Talk for a sample of findings from the LENA Natural Language Study.

## 2.0 METHODS

---

### 2.1 Demographics

A total of 334 children ages 2 months – 48 months from monolingual English-speaking households began the normative Phase I of the LENA Natural Language Study. Due to attrition and other factors, 311 participants completed this phase. There were 329 participants who contributed at least one valid 12-hour recording during Phase I.

### 2.2 Inclusion/Exclusion Criteria

To ensure a representative sample, we recruited children across an even age distribution and tried to match the US census with respect to the mothers' attained education level, which has been shown to be correlated with child language development (e.g., Arterberry, Midgett, Putnick, & Bornstein, 2007). We selected families according to the following criteria:

- 1) **Age distribution:** We recruited roughly eight children from each age month, ages 2 – 36 months, plus 15 children ages 37 – 48 months.
- 2) **Mother's education:** The normative development sample was intended to represent the US population with respect to mothers' education level (i.e., 23% college diploma, 29% some college, 26% high school diploma and no college, 22% no high school diploma). The goal was to match this distribution for each age month interval. For example, for the eight children who were six months old at the beginning of the study, on average two of them should have had mothers with college degrees, two should have had mothers with some college experience, two should have had mothers with high school diplomas (but no college classes), and two should have had mothers who did not graduate from high school.
- 3) **Household language:** Participants were selected from English-speaking households only.
- 4) **Normative language development:** Families with children who had been diagnosed with a language or developmental delay or disability were excluded from the original development study, as the data were collected for the purpose of establishing normative information about a typically developing population.

We actively sought participants from a representative distribution of children who were born premature, children who attended daycare, and children who had older or younger siblings, since any of these factors could influence language development.

Table 1 summarizes the distribution of the 329 participants who contributed at least one valid recording with respect to mothers' attained education levels compared to the US census. Appendix A lists the distributions of mothers' education levels, child gender and number of participants at each age level at the beginning of the normative Phase I of the LENA Natural Language Study (January, 2006).

**Table 1: Mother's Attained Education Compared To the 2004 US Census.<sup>2</sup>**

Mother's Education <sup>a</sup>	N	Normative Sample	US Census <sup>b</sup>
Some High School	45	14%	22%
High School Diploma	108	33%	26%
Some College	92	28%	29%
College Degree or Higher	84	25%	23%
Sample Total	329	100%	100%

<sup>a</sup> Some High School includes participants with no diploma; High School Diploma includes participants with a high school diploma, GED, or Trade School degree; Some College includes participants with some college courses but no bachelor's degree; and College Degree or Higher includes those with at least a bachelor's degree.

<sup>b</sup> US Census Bureau (2004): Population of women 15-44 years of age.

2 Note that the percentages in the first two education groups has changed since LTR-01-1 because the participants who had obtained GEDs were moved from the 'Some High School' group to the 'High School Diploma' group to more closely reflect the census grouping.

### 2.3 Attrition and Elimination

Of the 334 participants who started the study, 311 completed it. Ten participants proved to be too difficult to work with and were asked to leave the study, and thirteen participants dropped out because they moved out of the area or for other reasons.<sup>3</sup>

### 2.4 Participant Recruitment

We recruited participants through advertising in local newspapers and direct mail solicitation. Interested parents responded by contacting a call center representative who asked them to provide demographic information about their child and household. We selected potential participants based on criteria such as child age and mother's education level. Research assistants called to explain the study further; parents who remained interested were sent an informed consent form and questionnaire to collect additional demographic information. In total, 334 parents returned the consent form and were assigned participant ID numbers. Table 2 details the number of respondents at each stage of the recruitment process.<sup>4</sup>

**Table 2: Number of Respondents at Each Recruitment Stage.**

Stage of Recruitment	N
Responses to recruitment ad	1998
Potential participants selected	435
Potential participants reviewed consent form	364
Participants returned signed consent form	334

#### *Additional recruitment of 2-month-olds*

To ensure we collected a sufficient number of audio recordings for the normative database from very young children, we recruited approximately eight 2-month-olds each month between February-June, 2006. Those recruited in February contributed five recordings from February through June, those recruited in March contributed four recordings from March through June, and so forth. Note that for purposes of simplicity, the summary in Table 2 includes information for children recruited from February through June. Appendix B provides further information about the number of 2-month-olds recruited each month after January.

<sup>3</sup> Most of the 23 participants who dropped out or were eliminated provided no usable recordings.

<sup>4</sup> Of the 435 families who were invited to participate, 117 previously had participated in a pilot study conducted by the LENA Foundation in 2005.

## 2.5 Materials<sup>5</sup>

### *Preliminary documents*

Prior to the first recording session, participants were asked to sign an informed consent form detailing the study procedures and requirements for participation. They also completed a demographics questionnaire that requested additional detailed information about the child and the household.

### *Recording session materials*

Participating families received a packet containing recording materials the day before a scheduled recording session. The packet included several instructional documents. The How to Record Booklet provided step-by-step instructions about what to do on the recording day and quick reference inserts about materials and study protocol.

At the end of each recording session parents completed a Session Questions form to provide detailed information about the events related to the specific recording session. Parents of children who were between 8 months – 30 months of age were also asked to complete the MacArthur Communicative Development Inventory (Fenson et al., 2007), a parent self-report survey that asks about the child’s language development and the types of words that the child says/understands.

---

5 Samples of all materials are available on request.

**Professional evaluation session materials**

Table 3 describes the standard developmental assessments administered during professional evaluation sessions. Not all assessments were used during a session due to age or time constraints.

**Table 3: Standard Assessments Administered During Professional Evaluation Sessions.**

<b>Standardized Assessment</b>	<b>Type</b>	<b>Time (min)</b>
Receptive-Expressive Emergent Language Test-3 (REEL-3)	Parent Interview	20
Preschool Language Scale-4 (PLS-4)	Parent Interview/ Child Observation	20-30
Cognitive Adaptive Test/Clinical Linguistic and Auditory Milestone Scale (CAT/CLAMS)	Parent Interview/ Child Observation	6-20
Peabody Picture Vocabulary Test (PPVT)	Picture Pointing	12
Goldman-Fristoe Test of Articulation (GFTA)	Child Verbal Response	10
Bayley Scales of Infant and Toddler Development	Parent Interview/ Child Observation	40-60

## 2.6 Apparatus

Phase II participants recorded with an early prototype of the LENA digital language processor (DLP). This DLP prototype (see Figure 1) weighed 2.5 ounces. In order for the battery to run for a full day, the LENA DLP required charging for at least four hours using the LENA charger (Figure 1). Participants were sent LENA vests to wear on their recording days (Figure 1). The LENA DLP slipped into the front pocket of the LENA vest.



Figure 1. LENA DLP Prototype, Charger, and Vest Used by Study Participants

## 2.7 Design

### *Phase I: Recording Sessions*

All participants during Phase I were asked to record one day each month and were required to record for at least 12 consecutive hours each session. In addition, 61 participants completed a “double recording session” whereby they recorded on two consecutive days.

### *Phase I: Language Evaluation Sessions*

Nearly all participants visited the LENA Foundation at least once to be evaluated by a certified speech language pathologist (SLP) who administered between 3-4 standard language assessments.<sup>6</sup> The purpose was to obtain an independent assessment of each participating child’s language abilities.

Approximately half of the participants completed two additional evaluation sessions (one every two months) during the six month study. The purposes of the repeated sessions were 1) to determine the reliability of the individual assessments over time, and 2) to investigate the correlation between change over time in the assessments with change over time in the acoustic properties in the audio recordings. Table 4 shows the distribution of

<sup>6</sup> We were unable to schedule assessment sessions with 16 participants.

participants who completed single vs. repeated evaluation sessions by maternal education level. Table 5 lists average standard scores for the three most frequently administered language assessments.

**Table 4: Distribution of Participants Completing Single and Repeated Observation Sessions by Maternal Education Level.**

	Maternal Education Level <sup>a</sup>				Total
	Some High School	High School	Some College	College	
No Session	5	8	2	1	16
Single Session	16	51	41	37	145
Repeated Session	24	49	49	46	168
Total	45	108	92	84	329

<sup>a</sup> See Table 1 note for explanation of categories.

**Table 5: Average Standard Scores for Language Assessments.**

Scale	N	SS Mean	SS SD
REEL-3 Ability Score	263	102	13
PLS-4 Total Language Score	294	106	13
CAT/CLAMS Full DQ Score <sup>a</sup>	258	107	14

<sup>a</sup>Developmental Quotient (DQ) = (Developmental Age/Chronological Age)\*100

### *Phase I: Cognitive Evaluation Sessions*

We selected a subset of participants (N=79) for an additional cognitive evaluation. During these visits a trained professional research assistant administered the Bayley Scales of Infant and Toddler Development (Bayley, 2006), a standardized assessment that provides information about cognitive abilities. Cognitive evaluations took place once every two months for six months. The purpose of these evaluations was to obtain detailed information about participants’ intellectual abilities for future analyses and correlations with LENA

measurements. Table 6 details the demographic distribution of participants who completed cognitive evaluations by maternal education level.

**Table 6: Maternal Education Distribution for Participants Completing Cognitive Assessment Sessions.**

	Education Level <sup>a</sup>				Total
	Some High School	High School	Some College	College	
Number of Participants	9	25	24	21	79

<sup>a</sup> See Table 1 note for explanation of categories.

*Phase II: Extended Longitudinal Study*

Eighty Phase I participants were selected to participate in an extended longitudinal study and continue to provide natural language environment data. Phase II participants were chosen to provide a representative sample with respect to mother’s education. Participants continue to provide monthly audio recording data and to complete language evaluation sessions at approximately six month intervals. All valid recordings from this extended study are included in the normative database; thus, these participants may have contributed more than six recordings. Appendix C provides demographic information for Phase II (extended longitudinal) participants.

**2.8 Procedures**

*Recording sessions*

For the first two recording sessions in Phase I, parents were contacted individually to schedule recording session appointments. During the second phone call, research assistants scheduled additional appointments until the end of the study by assigning a “magic number” to each family (e.g., the 9<sup>th</sup> of the month). Parents were asked to record on the same day each month, corresponding to their magic number. This procedure conserved staff resources for scheduling time and also ensured that the recording sessions would be on different days of the week each month. Appendix D shows the distribution of Phase I and II recording sessions by day of the week.

Parents received a recording packet at least one day before their recording sessions via FedEx. They were instructed to take the charger out of the recording materials packet immediately and charge the LENA DLP overnight. We asked parents to begin recording as soon as their child woke up in the morning and to record continuously until their child went to bed that night. Parents were informed that should they be uncomfortable with some aspect of the recording session, it would be erased and not included in the normative data at their request and at no penalty to them.

Once activated, the LENA DLP could not be turned off. We told parents to remove the LENA vest during baths or nap time (the vest is not intended as sleepwear), but to place it near the child and to continue recording during that time. Participants were asked to behave as they would on any other day and to engage in any regularly scheduled routines with one exception: for the first three months of the study, parents were asked to turn off any ambient noise (e.g., TV, radio), and for the second three months of the study they were told that ambient noise was okay.

At the end of a recording session day parents were instructed to complete the included paperwork (i.e., Session Questions, language questionnaires) and to put all materials into a FedEx return envelope which they left on their doorstep for pick up the next morning. On delivery to the LENA Foundation, we uploaded the speech data to our electronic database and entered the other documentation into various databases.

### *Language and Cognitive Evaluation Sessions*

Evaluation session appointments were scheduled within two weeks of participants' recording sessions. Participants were the same age (in months) for each evaluation session and the corresponding recording session.

We asked participants to come to the LENA Foundation, Inc. in Boulder, Colorado for language and cognitive evaluations. They were told that the sessions would last approximately one hour and that during the session their child would interact with a LENA Foundation employee who would ask the child to do things like repeat words or point to pictures. Parents typically scheduled appointments for weekdays, but special Saturday evaluations were sometimes arranged. If a family had been selected for repeated evaluations but had scheduling conflicts, they were switched to single session status and completed only one evaluation. Parents were not permitted to bring siblings to appointments. At the start of a

language evaluation session, the SLP explained to parents that she would be interacting with the child and taking notes. She told parents the sessions would be videotaped to provide a visual record of each child's development.

The SLP administered as many assessments as was feasible during a one hour period or for as long as the child was deemed to be sufficiently attentive. Typically, the Receptive-Expressive Emergent Language Test, 3<sup>rd</sup> Ed (REEL-3) (Bzoch, League, & Brown, 2003), the Preschool Language Scale, 4<sup>th</sup> Ed (PLS-4) (Zimmerman, Lee, Steiner, & Pond, 2002), and the Cognitive Adaptive Test (CAT), and the Clinical Linguistic and Auditory Milestone Scale (CLAMS) (Accardo & Caput, 2005) were administered. If time permitted and the child was amenable, then the Peabody Picture Vocabulary Test (PPVT) (Dunn and Dunn, 1997) or the Goldman Fristoe 2, Test of Articulation (GFTA) (Goldman and Fristoe, 2000) were administered. After each evaluation, the SLP scored the assessments and entered the information into an electronic database. Participants who completed cognitive evaluations were administered the Cognitive, Gross Motor and Fine Motor sections of the Bayley Scales of Infant and Toddler Development, plus, time permitting, the expressive and receptive language sections. Research assistants entered all data a second time to provide a check on data entry errors.

### *Participant compensation*

Participants were compensated \$75 for each recording session (\$6.25/hour) and \$100 for each evaluation session (\$50 for the session, \$25 for travel and \$25 for child care). Participants had the opportunity to earn a \$200 bonus at the end of the study. We provided a list of bonus violations (e.g., missing evaluation appointments, forgetting recording session appointments, etc.) and deducted \$50 from the bonus for each violation. Most participants received some portion of the bonus.

## 3.0 RESULTS

### 3.1 Estimating Normative Population Values

A primary goal of the Phase I and Phase II studies was to determine population estimates for LENA measurements for reference purposes. All 3,066 recording sessions completed by participants from January, 2006 through December, 2007 were considered, and 87.5% ultimately were included in the normative set presented here.<sup>7</sup> We excluded 12.5% of the recording sessions as detailed in Table 7.<sup>8</sup> The final normative sample included 2,682 12-hour recording sessions from 329 participants ages 2 months to 48 months.

**Table 7: Normative Recording Sample Exclusions.**

	Count	Sample %
Original Sessions	3066	100.0
Exclusion Category		
Recording < 12 Hours	237	7.7
Age > 48 Months	68	2.2
DLP Error	42	1.4
Participant Error	10	0.3
Extreme Outlier Data	27	0.9
Final Included	2682	87.5

7 Although Phase II participants continue to provide monthly recording data, the analyses reported here include the first 18 months of Phase II data collected, from July, 2006 through December, 2007. Normative estimates will be updated with additional Phase II data periodically.

8 Counts presented here may vary from those in LTR-02-1, both due to more recent recordings being added and to some recording data previously excluded due to DLP error having been recovered.

### *Adult Word Count*

The Adult Word Count (AWC) report within the LENA System software estimates the total number of adult words the child hears per hour, per day and per month. Given that the range of ages of participants (2 months to 48 months) extends across developmental periods and that participants typically recorded over a six month (or longer) span, any significant increase or decrease in AWC with age could bias normative estimates. Thus, before estimating normative values, we examined the relationship between AWC and chronological age. We found no significant correlation ( $r(327)=.04$ ,  $p=.47$ ), and thus no age-related adjustments to AWC norms were made.<sup>9</sup>

During Phase I, participants contributed from one to seven recording sessions to the final normative sample ( $M=4.6$ ,  $SD=1.4$ ). During Phase II, participants contributed an additional 1 to 19 sessions ( $M=14.6$ ,  $SD=4.7$ ). To control for the variable number of recording sessions being contributed by each participant family, we first computed each family's average AWC from all usable sessions and from these values computed the full sample mean and standard deviation ( $M=12,297$ ,  $SD=6,462$ ).<sup>10</sup> We generated AWC estimates for the 1<sup>st</sup> to the 99<sup>th</sup> percentiles based on these values and assuming a standard normal (Gaussian) distribution. We implemented AWC estimates for each percentile into the LENA software as floor values; for example, the display for LENA System users with counts greater than or equal to the 50<sup>th</sup> but less than the 51<sup>st</sup> percentile shows a ranking at the 50<sup>th</sup> percentile.<sup>11</sup>

### *Conversational Turns*

The Conversational Turns (CT) report within the LENA System software provides an estimate of the total number of conversational turns the child engages in with an adult per hour, per day and per month. As might be expected and in contrast to AWC, CT counts increase significantly by month of age ( $r(327) = 0.51$ ,  $p<.01$ ); further analyses suggested that both the variability and degree of positive skew of CT counts also increase with age. Thus, we computed a unique mean and standard deviation for CT for each month of age.<sup>12</sup> Table 8 displays final CT mean and SD by age month.

---

9 AWCs for this analysis were totals from participants' first usable 12-hour recordings.

10 To adjust for skewing commonly seen with count data, we applied a square root transformation to normalize the distribution. Percentile values were estimated from these transformed data and then rescaled back to the original count metric.

11 We rounded AWC estimates to the nearest integer, constrained all AWC estimates to be  $\geq 0$ , and fixed the estimate for the first percentile to zero.

12 Each participant contributed at most one recording per age month.

**Table 8: Conversational Turns (CT) Estimates by Month of Age.<sup>a</sup>**

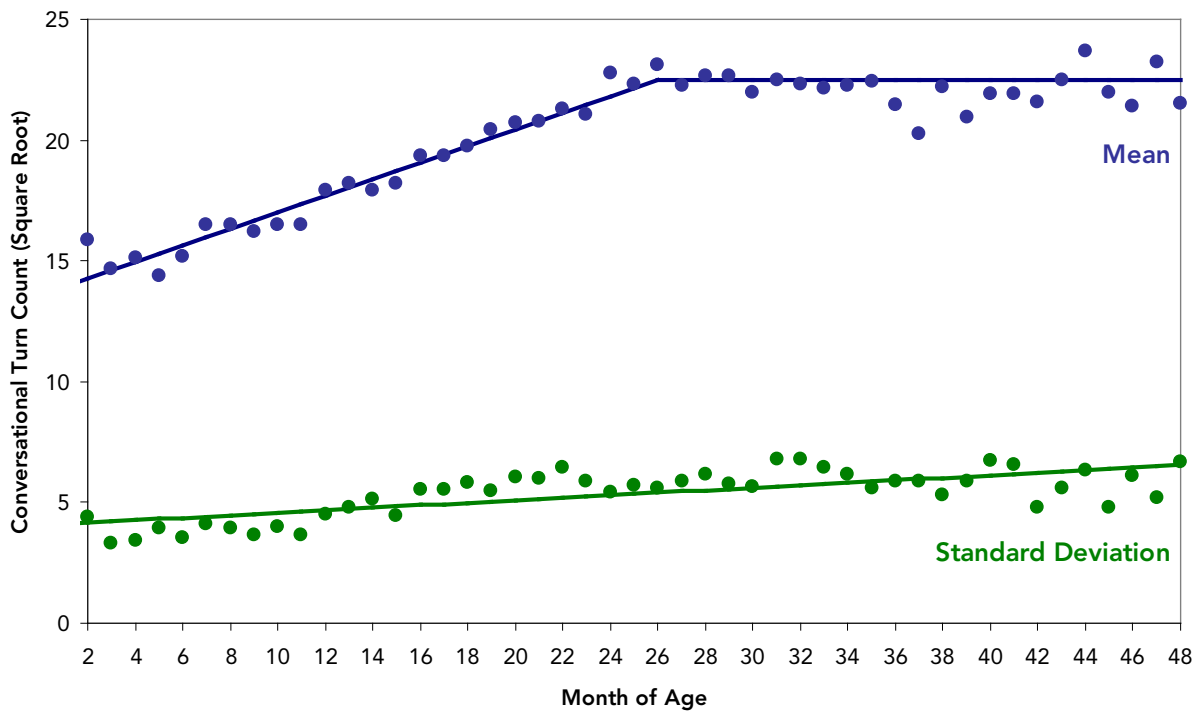
Age	N	Mean	SD	Age	N	Mean	SD	Age	N	Mean	SD	Age	N	Mean	SD
2	16	252	160	14	72	322	210	26	73	534	290	38	48	492	264
3	36	216	107	15	75	332	182	27	77	495	295	39	44	438	281
4	43	228	115	16	79	374	246	28	79	514	316	40	38	480	340
5	48	207	128	17	77	375	246	29	71	513	295	41	25	480	330
6	51	230	120	18	88	391	263	30	71	484	279	42	24	464	230
7	58	273	152	19	82	418	254	31	66	505	351	43	21	505	283
8	60	271	144	20	85	430	288	32	64	498	348	44	20	561	340
9	61	263	132	21	83	433	286	33	68	490	326	45	19	483	233
10	62	273	149	22	82	454	318	34	62	496	313	46	17	457	299
11	64	272	135	23	81	445	283	35	60	504	283	47	16	540	268
12	67	321	182	24	77	519	275	36	56	460	288	48	16	463	334
13	73	332	197	25	75	498	288	37	52	410	272				

<sup>a</sup> N is the total number of recordings/participants included at each age month.

To reduce age-related and random variability further and to improve interpretability, we performed separate regressions of CT means and standard deviations across month of age.<sup>13</sup> From the best-fit regression solutions we computed estimated CT means and standard deviations for each month of age from 0 months to 48 months. We generated final CT counts for the 1<sup>st</sup> to the 99<sup>th</sup> percentiles for each age month based on these values assuming a standard normal (Gaussian) distribution as was done for AWC.<sup>14</sup> Figure 2 displays CT mean and standard deviation values along with the best-fit regression lines.

13 CT means were fit using a two-segment linear spline with a knot at 26 months. CT standard deviations were fit via a 1<sup>st</sup>-order polynomial (linear) model.

14 See footnotes 10 & 11 (AWC calculations) for additional details.



**Figure 2. Best-fit Regression Solutions for Conversational Turn Counts.**

*Child Vocalization Frequency*

The Child Vocalization (CV) report within the LENA System software provides an estimate of the total number of vocalizations the child produces per hour, per day and per month. As was true for CT counts, CV counts increase significantly by month of age ( $r(327) = 0.61$ ,  $p < .01$ ), and further analyses suggested that both the variability and degree of positive skew of CV counts also increase with age. Thus, we computed a unique mean and standard deviation for CV for each month of age.<sup>15</sup> Table 9 displays final CV mean and SD by age month.

<sup>15</sup> Each participant contributed at most one recording per age month.

**Table 9: Child Vocalization (CV) Estimates by Month of Age.<sup>a</sup>**

Age	N	Mean	SD	Age	N	Mean	SD	Age	N	Mean	SD	Age	N	Mean	SD
2	16	736	368	14	72	1219	652	26	73	2160	931	38	48	2085	1195
3	36	726	389	15	75	1275	657	27	77	2052	897	39	44	2055	1217
4	43	837	390	16	79	1391	786	28	79	2132	1086	40	38	2181	1279
5	48	873	456	17	77	1503	834	29	71	2261	1186	41	25	2198	1248
6	51	929	491	18	88	1533	894	30	71	2206	1054	42	24	2231	935
7	58	1046	522	19	82	1611	786	31	66	2098	1099	43	21	2446	1138
8	60	1126	478	20	85	1738	900	32	64	2096	1236	44	20	2541	1194
9	61	1076	479	21	83	1736	887	33	68	2241	1326	45	19	2361	1012
10	62	1137	574	22	82	1794	1091	34	62	2186	1216	46	17	2384	1379
11	64	1141	515	23	81	1873	1083	35	60	2138	1127	47	16	2550	1367
12	67	1189	544	24	77	2152	938	36	56	2150	1299	48	16	2105	1172
13	73	1271	618	25	75	2115	982	37	52	1964	1346				

<sup>a</sup> N is the total number of recordings/participants included at each age month.

As was done for conversational turns, we performed separate regressions of CV means and standard deviations across month of age to improve interpretability and to reduce variability.<sup>16</sup> We generated estimated CT means and standard deviations for each month of age from 0 months to 48 months from these best-fit regression solutions. Again assuming a standard normal (Gaussian) distribution, we then generated final CV counts for the 1<sup>st</sup> to the 99<sup>th</sup> percentiles for each age month.<sup>17</sup> Figure 3 displays CV mean and standard deviation values along with the best-fit regression line.

16 CV means were fit using a two-segment linear spline with a knot at 26 months. CV standard deviations were fit via a 1<sup>st</sup>-order polynomial (linear) model.  
 17 See footnotes 10 & 11 (AWC calculations) for additional details.

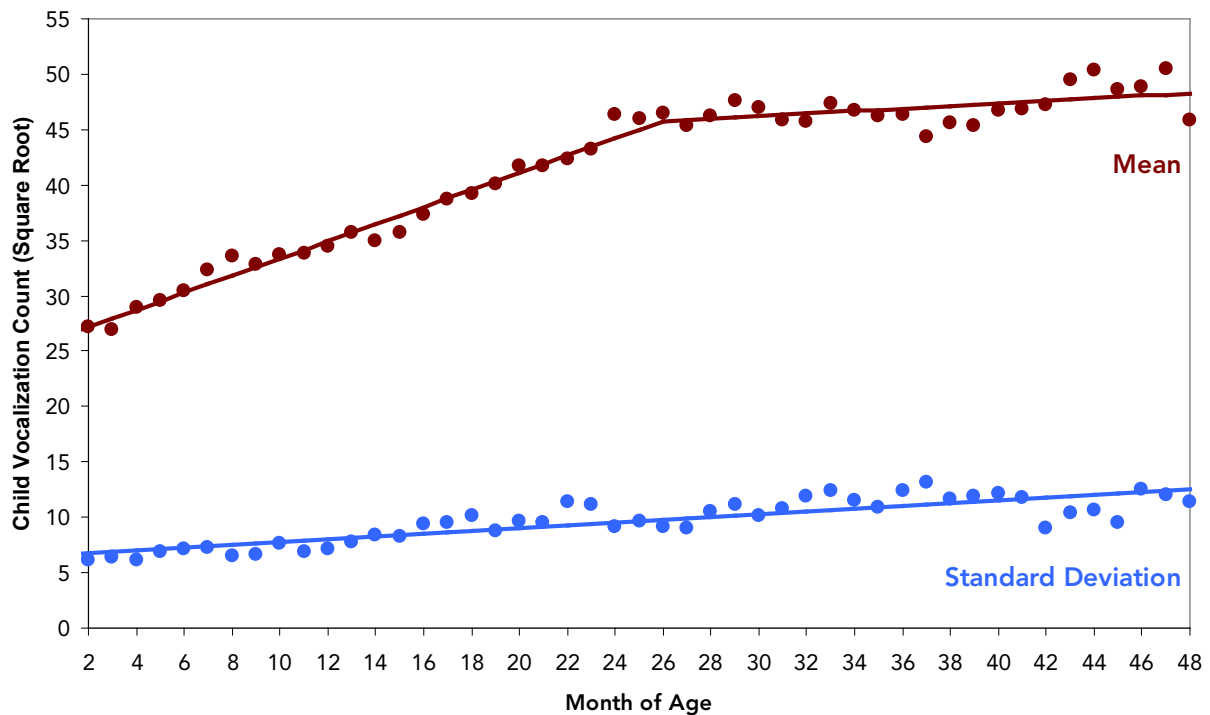


Figure 3. Best-fit Regression Solutions for Child Vocalization Counts.

### 3.2 Normative Values for Comparison Groups

We estimated AWC and CT values for a college-educated comparison group by selecting 696 audio recordings contributed by 84 families from the normative sample in which the mothers had graduated college. We computed the mean AWC for this subsample and determined that the average college-educated AWC fell near the 70<sup>th</sup> percentile. We then generated a new set of comparison data based on the subsample mean. Similarly, we examined data from 102 participants in a separate pilot user study and determined the average LENA software user’s AWC fell near the 76<sup>th</sup> percentile compared with the reference sample. Based on this value we generated a second set of comparison data.

## 4.0 SUMMARY

---

The research described in this paper represents the first known attempt to assemble and analyze a large corpus of full day spontaneous speech data from the households of infants and toddlers. Creating the first such database, this study establishes the baseline for future research. However, we must acknowledge that the lack of a valid external source of comparison data renders it difficult to identify unusual or suspect patterns of results. We hope that other researchers will use the LENA System to collect similar data that may further validate the results reported here and in the LENA software V3.1.0.

Although we made every effort to recruit a sample that was representative of the US population, there nevertheless remain several potential sources of selection bias. First, the sample population only included parents who responded to our recruitment ad. It is unclear whether these parents may have been more enthusiastic than other parents and perhaps might have had higher word counts or more advanced children. Second, although originally we included in our “No High School Education” group parents who had completed the GED, we have now moved them to the “High School” group to conform more closely to US Census grouping. This change makes more clear that the group with the least education is likely to be underrepresented in our sample.

Research at the LENA Foundation is ongoing, and our research participants continue to provide data that will enable us to refine our normative estimates further. In particular, as our sample of older children increases, the variability in our AWC, CV, and CT estimates for these ages should be reduced.

## REFERENCES

---

- Accardo, P.J. and Capute, A.J. *The Capute Scales: Cognitive Adaptive Test/Clinical Linguistic & Auditory Milestone Scale*. Baltimore: Paul H. Brookes Publishing Co., Inc., 2005.
- Arterberry, M.E., Midgett, C., Putnick, D.L., & Bornstein, M.H. (2007). Early attention and literary experiences predict adaptive communication. *First Language, 27*(2), 175-189.
- Bayley, N. *Bayley Scales of Infant and Toddler Development, Third Edition*. San Antonio: Harcourt Assessment, Inc., 2006.
- Bzoch, K.R., League, R., and Brown, V.L. *Receptive-Expressive Emergent Language Test, Third Edition*. Austin: PRO-ED, 2003.
- Dunn, L.M. and Dunn, L.M. *Peabody Picture Vocabulary Test, Third Edition*. Circle Pines: American Guidance Service, 1997.
- Fenson, L., Dale, P.S., Reznick, J.S., Thal, D., Bates, E., Hartung, J.P., Pethick, S., & Reilly, J.S. *MacArthur-Bates Communicative Development Inventories, Second Edition*. Baltimore: Paul H. Brookes Publishing Co., 2007.
- Goldman, R. and Fristoe, M. *Goldman Fristoe 2, Test of Articulation*. Circle Pines: American Guidance Service, Inc., 2000.
- Hart, B., and Risley, T.R. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore: Paul H. Brookes Publishing Co., Inc., 1995.
- Mullen, E.M. *Mullen Scales of Early Learning, AGS Edition*. Circle Pines: American Guidance Service, Inc., 1995.
- Zimmerman, I.L., Steiner, V.G., and Pond, R.E. *Preschool Language Scale, Fourth Edition*. San Antonio: The Psychological Corporation, 2002.

**Appendix A. Gender and Education Distribution by Age at First Recording.**

Age	Female	Male	Some High School	High School	Some College	College	Total
2	7	9	3	2	6	5	16
3	6	13	4	6	5	4	19
4	9	4	1	3	4	5	13
5	2	5	1	0	3	3	7
6	2	5	1	3	3	0	7
7	9	4	2	1	8	2	13
8	4	5	1	4	1	3	9
9	4	5	2	4	2	1	9
10	4	6	1	3	4	2	10
11	5	1	2	0	2	2	6
12	2	4	0	2	1	3	6
13	7	5	1	7	1	3	12
14	4	2	2	3	0	1	6
15	7	2	1	6	1	1	9
16	5	6	1	4	2	4	11
17	3	2	1	2	2	0	5
18	4	7	2	4	2	3	11
19	2	3	0	1	3	1	5
20	6	6	2	6	3	1	12
21	3	3	1	2	2	1	6
22	6	1	1	1	1	4	7
23	5	6	2	5	3	1	11
24	4	4	0	3	3	2	8

**Appendix A. Gender and Education Distribution by Age at First Recording. (cont.)**

Age	Female	Male	Some High School	High School	Some College	College	Total
25	4	2	0	3	3	0	6
26	5	4	1	4	1	3	9
27	7	4	2	3	3	3	11
28	6	4	1	2	4	3	10
29	1	5	0	4	0	2	6
30	2	5	2	2	3	0	7
31	4	5	1	1	3	4	9
32	4	4	2	2	3	1	8
33	3	4	2	3	1	1	7
34	3	5	0	4	2	2	8
35	3	5	0	1	3	4	8
36	3	4	2	3	1	1	7
37	3	2	0	1	3	1	5
38	0	0	0	0	0	0	0
39	2	2	0	2	0	2	4
40	0	0	0	0	0	0	0
41	0	2	0	0	0	2	2
42	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0
44	2	1	0	1	0	2	3
45	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0
47	0	1	0	0	0	1	1
48	0	0	0	0	0	0	0
<b>Total</b>	<b>162</b>	<b>167</b>	<b>45</b>	<b>108</b>	<b>92</b>	<b>84</b>	<b>329</b>

**Appendix B. Gender and Education Distribution for 2-Month-Olds.**

Month	Female	Male	Some High School	High School	Some College	College	Total
Feb	5	5	2	2	3	3	10
Mar	3	4	1	1	3	2	7
Apr	2	5	1	4	1	1	7
May	3	3	0	0	3	3	6
Jun	1	1	0	0	2	0	2
<b>Total</b>	14	18	4	7	12	9	32

**Appendix C. Gender and Education Distribution by Age For 80 Longitudinal Participants in July 2006.**

Age	Female	Male	Some High School	High School	Some College	College	Total
3	1	1	0	0	2	0	2
4	1	0	0	0	1	0	1
5	2	2	1	0	2	1	4
6	3	3	1	1	2	2	6
7	1	0	0	0	1	0	1
8	2	0	1	1	0	0	2
9	0	3	0	2	0	1	3
10	2	2	1	0	2	1	4
11	2	1	0	2	0	1	3
12	2	2	0	1	3	0	4
13	2	0	0	0	2	0	2
14	2	2	0	0	1	3	4
15	0	1	0	0	1	0	1
16	2	1	0	1	2	0	3
17	1	1	1	0	0	1	2
18	2	0	0	0	0	2	2
19	1	1	0	1	0	1	2
20	1	0	0	1	0	0	1

**Appendix C. Gender and Education Distribution by Age For 80 Longitudinal Participants in July 2006 (cont.).**

Age	Female	Male	Some High School	High School	Some College	College	Total
21	1	0	0	1	0	0	1
22	1	3	0	2	1	1	4
24	0	1	1	0	0	0	1
25	1	0	1	0	0	0	1
26	1	2	0	1	2	0	3
27	1	0	0	0	1	0	1
28	1	0	0	0	0	1	1
29	2	0	0	1	1	0	2
30	0	1	0	1	0	0	1
31	0	1	0	1	0	0	1
32	2	0	0	0	1	1	2
33	1	0	1	0	0	0	1
34	1	1	1	1	0	0	2
36	1	2	1	1	0	1	3
37	0	4	1	1	2	0	4
38	2	0	0	0	1	1	2
39	1	1	1	0	1	0	2
46	0	1	0	1	0	0	1
<b>Total</b>	43	37	12	21	29	18	80

**Appendix D. Phase I & II Recording Sessions Per Day of the Week.**

<b>Day of the Week</b>	<b>Number of Recordings</b>	<b>% of Total</b>
Sunday	354	13.2
Monday	408	15.2
Tuesday	380	14.2
Wednesday	390	14.5
Thursday	396	14.8
Friday	412	15.4
Saturday	342	12.8
<b>Total</b>	<b>2682</b>	<b>100.0%</b>