Behav Res (2018) 50:1921–1932
DOI 10.3758/s13428-017-0960-0

Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings

Tobias Busch[1,2] · Anouk Sangen[1] · Filiep Vanpoucke[2] · Astrid van Wieringen[1]

Published online: 21 September 2017

Abstract The Language ENvironment Analysis system (LENA™) automatically analyzes the natural sound environments of children. Among other things, it estimates the amounts of adult words (AWC), child vocalizations (CV), conversational turns (CT), and electronic media (TV) that a child is exposed to. To assess LENA's reliability, we compared it to manual transcription. Specifically, we calculated the correlation and agreement between the LENA estimates and manual counts for 48 five-min audio samples. These samples were selected from eight day-long recordings of six Dutch-speaking children (ages 2–5). The correlations were strong for AWC, $r = .87$, and CV, $r = .77$, and comparatively low for CT, $r = .52$, and TV, $r = .50$. However, the agreement analysis revealed a constant bias in AWC counts, and proportional biases for CV and CT (i.e., the bias varied with the values for CV and CT). Agreement for detecting electronic media was poor. Moreover, the limits of agreement were wide for all four metrics. That is, the differences between LENA and the manual transcriptions for individual audio samples varied widely around the mean difference. This variation could indicate that LENA was affected by differences between the samples that did not equally affect the human transcribers. The disagreements and biases cast doubt on the comparability of LENA measurements across families and time, which is crucial for using LENA in research. Our sample is too small to conclude within which limits LENA's measurements are comparable, but it seems advisable to be cautious of factors that could systematically bias LENA's performance and thereby create confounds.

Keywords Automatic speech recognition · Human transcription · Dutch · Reliability · Agreement · Method comparison · Bias · Measurement error · Child vocalization count · Adult word count · Conversational turn count · Electronic media

The Language ENvironment Analysis system (LENA™; Xu, Yapanel, Gray, Gilkerson, et al., 2008) can automatically analyze the natural sound environments of children. Since it provides metrics that are more difficult to obtain through traditional observation and transcription, LENA has been well received by researchers. Yet, relatively little attention has been paid to LENA's reliability and its limitations as a measurement instrument.

Among other things, LENA has been used to investigate links between the language environment and language development (Weisleder & Fernald, 2013), to describe the characteristics of particular child care environments and caregivers (Caskey & Vohr, 2013; Johnson, Caskey, Rand, Tucker, & Vohr, 2014; Soderstrom & Wittebolle, 2013), to gain insights into the auditory environment and language development of clinical populations (e.g., Ambrose, VanDam, & Moeller, 2014; Oller et al., 2010; Thiemann-Bourque, Warren, Brady, Gilkerson, & Richards, 2014; Warlaumont, Richards, Gilkerson, & Oller, 2014), and as feedback in parent-centered interventions (Pae et al., 2016; Suskind et al., 2013; Zhang et al., 2015).

The validity of such studies hinges on the reliability of the LENA metrics—that is, the degree to which they reflect true variation rather than measurement error. When LENA scores are compared across families or points in time, the measurement error must not mask or distort relevant differences and should be unaffected by irrelevant differences.

✉ Tobias Busch
tbusch@cochlear.com

[1] Department of Neuroscience, Research Group ExpORL, KU Leuven, Leuven, Belgium

[2] Cochlear Technology Centre, Schaliënhoevedreef 20i, Mechelen, Belgium

The LENA system has multiple parts that could introduce error, and that thus require validation: It consists of a small recorder that is worn by the child for a day and software to later process the recording. The software runs a signal-processing algorithm that has been trained on a large sample of manually annotated natural language recordings (Gilkerson, Coulter, & Richards, 2009; Gilkerson & Richards, 2009) and uses acoustical features of the audio signal to estimate various speech and audio variables.

The LENA algorithm processes recordings in multiple steps: First, it extracts a number of acoustical features and uses them to partition the audio stream. Then each segment is assigned to a sound class (e.g., adult speech, child speech, silence). This labeling is done through maximum likelihood based Gaussian Mixture and Hidden Markov models. The speech segments are processed further—for example, to distinguish between speech and nonspeech sounds (such as cries, burps) produced by the child, and to estimate the number of adult words (Xu, Yapanel, Gray, & Baer, 2008b; Xu, Yapanel, Gray, Gilkerson, et al., 2008). Eventually, the automatic annotations are aggregated into various counts. Often-used counts are the number of vocalizations produced by the key child (i.e., the child wearing LENA; CV), adult words spoken to the child (AWC), conversational turns (CT), and audible electronic media (TV).

One way to assess the reliability of LENA is to compare it to manual transcriptions. Although manual transcriptions are themselves not entirely free of measurement error, they certainly are the gold standard for what LENA is trying to achieve. Comparisons between LENA and manual transcription can be based on either the algorithm's intermediate steps (the labels) or its final output (the counts).

A number of studies have compared the labels that LENA assigns to audio segments to those assigned by humans: Xu, Yapanel, and Gray (2009) transcribed 70 h of recordings from English-speaking American families and compared manual and LENA labels per 10-ms time window. LENA's sensitivity (i.e., how many human-labeled segments were assigned the same label by LENA) was high for adult speech (82%), child speech (76%), TV (71%), and "other" segments (76%). LENA also distinguished well between speech and nonspeech child vocalizations (e.g., crying, burping), with 75% and 84% accuracy, respectively. Another 12 h were added to this data set later, with comparable results (Xu, Richards, & Gilkerson, 2014).

Gilkerson et al. (2015) used the same method with 5.5 h of recordings from Chinese families. They found a similar sensitivity for Chinese adult and child speech (79% and 81%, respectively). Again, LENA distinguished speech and nonspeech child vocalizations well (84% and 70% sensitivity, respectively). However, contrary to Xu et al. (2009), many of LENA's child speech labels in this case were false positives (73%), such as misclassified motherese. Moreover, the sensitivity for TV was low (2%), with almost half of the TV segments being labeled as child or adult speech.

Both Xu et al. (2009) and Gilkerson et al. (2015) found that LENA classified many speech segments as "overlapping speech" or "other" (a catchall category that includes segments with low signal-to-noise ratios). Xu et al. (2009) showed that such misses might be more common in certain situations (e.g., outdoors). LENA's AWC and CV counts exclude segments with low signal-to-noise ratio. Interestingly, the threshold for exclusion seems to be rather low: Oller et al. (2010, supplemental material), using the data of Xu et al. (2009), found that when speech segments with low signal-to-noise ratios were included, LENA could still identify 73% of the child speech segments, with only 5% false positives.

VanDam and Silbert (2016) used a different method to evaluate LENA's labels: They let 23 trained judges categorize 2,340 audio snippets that LENA had labeled as either female adult, male adult, or key child speech. The mean Cohen's Kappa between individual judges and LENA was .68. Fleiss's Kappa for the overall agreement was .79.

In general, LENA and human transcribers seem to agree relatively well when labeling speaker and sound types. However, these labels relate only indirectly to the LENA counts of adult words, child vocalizations, conversational turns, and electronic media. Accurate labels are a prerequisite for accurate counts. Yet it is difficult to foresee, for instance, how the sensitivity of detecting adult speech or the exclusion of speech segments with low signal-to-noise ratio would affect AWC or CT counts.

Some studies have assessed the reliability of LENA's counts more directly through correlations with manual counts: For AWC, Xu et al. (2009) reported a correlation of $r = .92$ ($n = 70, p < .01$). Similarly, Oetting, Hartfield, and Pruitt (2009) found a correlation of $r = .85$ ($p < .001$) in interactions between 17 African American children and their mothers (notably, they played previously made recordings to the LENA recorder to obtain the estimates; somewhat surprisingly, the correlation for AWC changed to $r = .76$, $p < .001$, when LENA processed the recordings a second time). The LENA and manual AWC scores also correlated well for Spanish as spoken in the United States ($r = .8$, $n = 120$; Weisleder & Fernald, 2013, supporting material), Chinese ($r = .73$, $n = 22$, $p < .001$; Gilkerson et al., 2015), Korean ($r = .72$, $n = 63$, $p < .001$; Pae et al., 2016), and European French ($r = .64$, $n = 324$, $p < .001$; Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2016). Both Xu et al. (2009) and Canault et al. (2016) reported that LENA's AWC counts were on average lower than the manual counts.

Canault et al. (2016) is the only study that has investigated LENA's CV counts. Similar to AWC, the CV counts were lower than the manual counts, yet the correlation between them was high ($r = .71$, $p < .001$). Further analysis revealed an effect of the child's age and the signal-to-noise ratio on both AWC and CV counts.

Correlations between LENA and manual CT counts have only been reported by three of these studies: Oetting et al. (2009) found no correlation ($r = .14$, $n = 17$, $p > .05$, and $r = .08$, $p > .05$, on the second playback). Gilkerson et al. (2015) found significant correlations only after removing three outlying samples ($r = .72$, $n = 19$, $p < .001$). Similarly, Pae et al. (2016) found no significant correlation ($r = -.03$, $n = 63$, $p > .05$) unless five samples that contained much overlapping speech and whining were removed ($r = .67$, $n = 58$, $p = .001$).

Berends (2015) reports that for Dutch the LENA CT counts were significantly lower than manual counts (Kolmogorov–Smirnov test, $Z = 2.56$, $n = 14$, $p < .001$). Notably, 48% of the missed turns were caused by LENA misclassifying adult speech as child speech or vice versa, and 35% because LENA had labeled speech that was intelligible to the transcribers as noisy.

It should be noted that the samples used by these studies differed substantially, particularly regarding the sampling procedure: Some selected periods with high speech activity (Berends, 2015; Canault et al., 2016; Gilkerson et al., 2015; Xu et al., 2009), and others sampled equally from the entire duration of the recording (VanDam & Silbert, 2016) or along the entire distribution of AWC (Weisleder & Fernald, 2013). Unlike others, Canault et al. also made sure to sample different activities and environments. Oetting et al. and Pae et al., on the other hand, used recordings from rather controlled environments. Moreover, where most studies have used transcribed, noncontinuous 5- or 10-min segments, Oetting et al. used 30-min-long recordings.

After reviewing the literature, further investigations of LENA's reliability seem necessary, for two reasons: First, the results have varied between studies using different languages. This could be due to methodological differences or because the algorithm (which has been optimized for American English) relies on language-specific features. We were planning to use LENA with Dutch. Since an effect of language could not be ruled out and only the reliability of LENA's CT counts has been investigated for Dutch (Berends, 2015), a more comprehensive assessment seemed indispensable.

Second, the previous studies do not provide a complete picture of the disagreements and biases between manual transcription and LENA, especially regarding the aggregated counts we were most interested in—namely, the estimated amounts of adult words, child vocalizations, conversational turns, and audible electronic media. As we mentioned above, the agreement between LENA and manual labels relates only indirectly to these counts. Correlations, on the other hand, are unsuited to assess the agreement between LENA and manual transcriptions. Correlations between two methods are a means to assess concurrent validity. Insofar as human scores can be considered good approximations of the true score, they also reflect LENA's measurement error, and thus its reliability. Nevertheless, correlations only indicate the strength of linear association, not agreement.

Agreement is the degree to which the measurements are identical. As was pointed out by Ludbrook (1997), Pearson's product-moment correlation coefficient indicates the scatter of values around the line of best fit—it does not matter whether the slope of that line is different from 1 (proportional bias) or whether the intercept differs from 0 (fixed bias). Hence, correlation quantifies random error, but not the systematic biases between methods. Relying solely on correlations to assess whether one method can be replaced by the other can be misleading (Bland & Altman, 1986). Regressing the results of one method on those of another using ordinary least squares has also been criticized (Altman & Bland, 1983; Ludbrook, 1997).

A means to assess agreement between two methods has been proposed by Bland and Altman (1986; see also Bland & Altman, 1999; Carstensen, 2010; Giavarina, 2015). Their *limits of agreement* (LoA) approach focuses on the random and systematic differences between measurements made with two methods in parallel. It involves determining prediction intervals for the differences (the LoA) and identifying biases—that is, whether one method consistently produces higher values than the other (fixed bias), or whether the differences change with the to-be-measured quantity (proportional bias). The LoA approach aims to highlight biases and disagreements in order to provide insights beyond those from correlation coefficients or scatterplots. Notably, there is no fixed threshold above which agreement would be considered acceptable; whether the LoA are too wide or the biases too strong has to be decided for each application.

The main objective of the present study was to get a comprehensive picture of how reliably LENA measures the natural language environment of Dutch-speaking children. To do this, we compared the counts produced by LENA to those obtained from manual transcription, whereby manual transcription was considered the gold standard for what LENA was trying to achieve.

We focused on four LENA counts: AWC, CV, CT, and TV. Lena's CV count is based on acoustical criteria that we found difficult to implement manually. Thus, we instead compared it to the transcriber's count of "child utterances." The latter are based on semantic boundaries and often are used in language acquisition research (e.g., MacWhinney, 2000). To compare the two methods, we used correlations as well as the LoA approach.

The goal was by no means to approve or dismiss LENA on the basis of our results. Instead, we wanted to document its relation to manual transcriptions, highlight potential pitfalls for study design, and stimulate further discussion and investigation. Specifically, we wanted to know

1. How well does LENA correlate with manual transcriptions? How do our results for Dutch compare to those for other languages?

2. How well do LENA measures agree with manual transcription? Are there fixed and/or proportional biases?
3. Could some biases and disagreements become problematic when using LENA in a research study—that is, when comparing LENA measurements across subjects or points in time?

Although the literature indicates cross-linguistic variability, we expected the correlations for Dutch to resemble the rather high correlations reported for American English (which the LENA algorithm has been trained on). Both are West-Germanic languages, and thus similar. The vowel system of Dutch, for instance, is similar to that of English in many respects (Wang, 2007, chap. 3).

We also expected to find substantial disagreements and biases between the manual and LENA counts. Human transcribers can utilize contextual knowledge and linguistic insights. This could mean that they are less affected by such factors as noise, overlapping speech, and distance between speakers. In a diverse sample of more and less complex naturalistic recordings, this could lead to a large variation of the differences between the two methods (i.e., wide LoA), and even systematic biases. Moreover, our transcription protocol was not designed to mimic LENA's algorithm. Instead, we assessed the same key components of verbal interactions in the same way we would if LENA were not available. Overall, this comparison certainly set the bar high for LENA, and we were curious to see how it would perform.

Method

Participants

Six Dutch-speaking children (ages 2–5 years; two female, four male) from Flanders (Belgium) wore the LENA recorder for at least eight consecutive hours during regular days; two of the children made two day-long recordings. It should be noted that LENA's algorithm was trained on recordings from slightly younger children (2–48 months old; see Oller et al., 2010, supplementary material, Table S2). The age range in the present study was based on the age range with which we wanted to use LENA in the future.

The recordings were processed with LENA Pro version 3.4.0. In total, 1,084 five-min-long audio samples with corresponding LENA counts were obtained, and 48 of them were selected for manual transcription as described below.

Procedure

When comparing measurement methods, the sample should reflect the range over which the methods are used (Carstensen, 2010, p. 2), and the variance of the measured quantities should exceed the expected measurement error (Carstensen, 2010, p. 27). The main purpose of these restrictions is to increase the sensitivity to biases. Thus, it was important to assure that a wide range of all four quantities (adult words, child vocalizations, conversational turns and electronic media) would be represented in the transcribed sample.

To achieve this, we selected the samples for transcription as follows: All 1,084 five-min audio samples were sorted by their LENA AWC (which ranged from 0 to 827) and divided into 20 groups along the entire range (i.e., in steps of 50). A similar partitioning was done for the LENA counts of CV (range 0–153; 16 groups in steps of 10), CT (range 0–30; 15 groups in steps of 2), and TV (range 0–300; 10 groups in steps of 30).

The goal was to randomly select at least one 5-min audio sample from each group of each variable, so that the entire range of all four variables would be covered. Since all 5-min samples contained some amounts of adult speech, child's speech, conversational turns, and electronic media, each of them was included in four groups. Thus, picking one sample always meant drawing from four groups. Samples were drawn one at a time, and all previously selected samples were taken into account when deciding which groups still needed to be drawn from. Despite this, drawing multiple times from the same group was often unavoidable. As a result, the distribution of the four LENA variables in the 48 selected audio samples was not uniform.

The 48 selected samples were then transcribed by one of two native Dutch speakers to obtain the manual counterparts to LENA's AWC, CV, CT, and TV counts. To assess the interrater reliability of the transcription protocol, a subset of 17 out of the 48 audio samples were randomly selected and transcribed by both transcribers. This subset contained between two and five audio samples from each of six different recordings. The recordings came from five of the children in our sample (ages 2–5; one female, four male). One child could not be included due to a delay in the data collection.

Transcription protocol

The transcribers listened to each 5-min audio file without any knowledge of the corresponding LENA output. They used TextGrids in Praat (Boersma & Weenink, 2016) to make the annotations. In the TextGrids, they marked intervals of speech from each individual speakers and sounds from electronic media on separate interval tiers. The speaker (key child, mother, father, TV, etc.) was identified in the tier name. The transcribers used a point tier to mark conversational turns.

After all speech intervals had been marked, all speech from the key child and all speech from adults was transcribed. The transcribers used different tags to mark words of uncertain lexical status—that is, onomatopoeias, babbling, family-specific word forms (e.g., *bebe* for bear), neologisms, laughing, crying, and filled pauses (e.g., *uh*, *um*). They also indicated which words overlapped with other speaker's speech, and used a special notation to indicate overlapping speech that may have been understood by the child—namely, if the overlap was faint (e.g., another speaker in the background), very brief (e.g., at a conversational turn), or when child and parent were singing along.

When marking conversational turns, all conversations between the key child and any adult were marked—that is, all segments of adult speech followed by a response from the key child, or vice versa. As with the LENA algorithm, the first utterance of each conversation was not counted as a turn, and pauses of 5 s or more constituted the end of a conversation (see Ford, Baer, Xu, Yapanel, & Gray, 2008). Intermittent backchannels—such as *mhm, ja* [yeah] and *echt?* [really?]—are often used to signal engaged listenership; they arguably do not qualify as full turns and were not marked by the transcribers. It should be noted that LENA lacks the semantic insight to make this kind of distinction and might count backchannels as conversational turns.

The finished transcriptions were parsed with Python 3.5, using the TextGridTools library (Buschmeier & Włodarczak, 2013) to extract the manual AWC, CV, CT, and TV counts for each 5-min sample. The four counts were obtained as follows:

For AWC, we counted all Dutch words spoken by adults that the transcribers could understand or for which they inferred from the child's response that it had understood them. Common Dutch contractions were counted as single words—for example, *zo'n [zo een]*, *'tis [het is]*, and *da's [dat is]*. We also counted onomatopoeias, family-specific word forms (e.g., *bebe* for bear), and neologisms. Although their lexical status is uncertain, they arguably are lexical from the talker's perspective. Moreover, they are indistinguishable from regular words for the LENA algorithm and will be counted by it. Laughing, crying, filled pauses (e.g., *uh*, *um*), and babbling, on the other hand, were not counted as adult words. Words overlapped by speech from other speakers were also not counted, unless the transcribers indicated that the overlapping speech had likely been understood by the child (see above).

The concept underlying LENA's count of child vocalizations (CV) is that of "breath groups" (Oller et al., 2010, supplemental material, p. 17), whereby a 300-ms pause ends a vocalization and vegetative sounds (e.g., cries, burps, raspberries) are not counted. This allows LENA to deal with the high variability of speech in early childhood despite its lack of lexical and semantic insight. On the other hand, this is an unusual way to quantify child language production and difficult to implement reliably in a manual transcription protocol. We therefore chose an operationalization more similar to the concept of "utterances" as implemented in the popular Codes for the Human Analysis of Transcripts (CHAT)transcription system (MacWhinney, 2000). That is, vocalization boundaries were aligned with semantic boundaries. Hence, pauses in the middle of words or sentences did not constitute a boundary. As with the AWC count. Family-specific word forms and neologisms were treated as valid words (as they would be by LENA), whereas segments of child language production were not counted if they only contained vegetative sounds or other noncommunicative sounds (e.g., *hmm*, *uh*). However, child utterances that only contained babbling were counted because for younger children babbling constitutes relevant speech and because the LENA algorithm is going to count babbling as well. In our sample less than 1% of the child utterances contained nothing but babbling. This is no surprise, considering that all children were past the canonical babbling phase. LENA also identifies vegetative sounds like cries and laughs and does not count them, whereas noncommunicative sounds that are harder to distinguish from regular speech (including babbling) would likely be counted by LENA.

The CT count was simply the number of turns marked by the transcribers, and the TV count (i.e., the duration of audible electronic media) was the total duration of all electronic media intervals that the transcriber had marked (i.e., up to 300 s for a 5-min audio sample).

Statistical analysis

Statistical analyses were conducted using R version 3.2.5 (R Development Core Team, 2016). The interrater reliability of our manual transcription protocol was assessed with a subset of 17 five-min samples. We used the psych package in R (Revelle, 2016) to compute intraclass correlation coefficients (ICC), specifically a Sample × Transcriber two-way random analysis of variance with single measures [ICC(2,1); Shrout & Fleiss, 1979]. The ICC indicates the proportion of the variance that is attributable to differences between audio samples rather than between transcriptions. Thus, an ICC close to 1 means high agreement between the two transcribers.

For all four variables, Wilcoxon matched-pairs signed rank tests were used to determine whether the LENA counts were significantly different than the manual counts. We also calculated Pearson correlation coefficients and the LoA. The LoA are 95% prediction intervals for the differences between the methods—that is, they indicate the intervals in which 95% of the differences between measurements made by the two methods are expected to fall, assuming they are normally distributed.

To calculate LoA, we followed the guidelines of Carstensen (2010, chap. 4): For all 5-min samples mean and difference of LENA and manual counts were calculated. To identify fixed bias, we tested whether the mean difference was significantly different from zero using a one-sample *t* test. To identify proportional bias (i.e., nonconstant mean difference), the differences were regressed on the means and the slope of the regression line ($b_{1\text{ , diff}}$) was tested for statistical significance. To identify nonconstant variance of the differences (heteroscedasticity), the absolute residuals of that model were regressed on the means and the slope of the regression line ($b_{1,\text{res}}$) was tested for significance. For all tests, the significance level was set to $\alpha = .05$.

LoA are calculated as 95% prediction intervals around the mean difference between LENA and manual counts—that is, LoA = $M_{\text{diff}} \pm t_{.975,n-1} \times SD_{\text{diff}}$. In the simplest case of no proportional bias and constant variance, M_{diff} and SD_{diff} are estimated from the mean and standard deviation of the differences in the sample. If there is proportional bias, M_{diff} changes with the magnitude of the measured variable; their relation is estimated from the regression of the differences of LENA and human measures on their means (m) as $M_{\text{diff}}(m) = b_{0,\text{diff}} + m^{*}b_{1,\text{diff}}$. When there is nonconstant variance, SD_{diff} depends on the magnitude of the measured variable; their relation is estimated from a regression of the absolute residuals on the means as $SD_{diff}(m) = \left(b_{0,\text{res}} + m^{*}b_{1,\text{res}}\right) \times \sqrt{\pi/2}$.

Results

Descriptive statistics of the manual counts, the LENA counts, and the differences between them for the n = 48 transcribed samples can be found in Table 1. Scatterplots of the counts and histograms of the differences are shown in Fig. 1.

Interrater reliability

A subset of 17 randomly selected 5-min audio samples was transcribed by both transcribers to assess the interrater reliability. For AWC, the mean difference between the transcribers was 35.9 (SD = 88.1). The ICC was .93 with a 95% CI of [.82, .98]. For one sample the difference between the transcribers' AWC counts was 3.4 SD s above the mean, because of a background conversation only one of them had transcribed. Without this outlier, the ICC for AWC increased to .97 , 95 % CI [.75, .99].

The interrater reliability was also high for CV and CT, with a mean difference of 8.4 (SD = 7.6) and ICC = .92 , 95 % CI [.39, .98], for CV, and a mean difference of 4.3 (SD = 4.3) and ICC = .97 , 95 % CI [.75, .99], for CT.

Calculating ICC was not sensible for TV, since for all 13 samples both transcribers agreed that there was no TV. For two samples, the TV counts were almost identical (<1-s difference), and for the remaining two they differed by 17 and 22 s (7% and 12%, respectively)

Adult word count (AWC)

On average, LENA counted fewer adult words than the transcribers, $M_{\text{diff}} = -55.8$, $SD_{\text{diff}} = 125.3$. A Wilcoxon signed rank test indicated that the LENA AWC (Mdn = 118.5) were significantly lower than the human AWC (Mdn = 212), V = 821 , p = . 007. There was a strong correlation between LENA and manual values of AWC, $r(46) = .87$, $p < .001$.

We then calculated LoA—that is, the 95% prediction intervals for the differences between the methods. The regression of the differences on the means did not indicate a significant proportional bias, ($b_{1,\text{diff}} = -0.1$, $t(46) = -1.3$, $p = .22$), but the regression of the absolute residuals on the means revealed nonconstant variance ($b_{1,\text{res}} = 0.13$, $t(46) = 2.63$, $p = .012$). Thus we estimated the LoA, whose width changes with the number of adult words in the recording m, as $LoA(m) = -31{-}0.1m \pm 2.01 \times (53.4 + 0.13m)\sqrt{\pi/2}$.

We used this model of LoA to predict the expected mean difference and LoA for 5-min samples with different numbers of adult words. Across all 1,084 five-min samples, the median LENA AWC count was 51. For a sample containing 51 adult words, the predicted mean difference between LENA and human is LoA(51) = − 36, and 95% of the differences are expected to fall within ± 152 words around that (i.e., between – 188 and + 116 words). As the number of words increases, the LoA become wider: For samples with 141 adult words (the 75th percentile of the LENA AWC counts in the full sample), the predicted mean difference is – 44 and the LoA is ± 182 words around that (i.e., – 226 to + 137 words; Fig. 1).

Child vocalizations (CV)

The mean difference between the LENA and manual CV counts was $M_{\text{diff}} = 0.92$ ($SD_{\text{diff}} = 25.6$). A Wilcoxon signed rank

Table 1. Descriptive statistics of manual counts, LENA counts, and differences between the manual and LENA counts for n = 48 transcribed 5-min segments

Count	Lena				Manual				Difference (Lena – Manual)			
	M (*SD*)		*Mdn* [Q1; Q3]		*M* (*SD*)		*Mdn* [Q1; Q3]		*M* (*SD*)		*Mdn* [Q1; Q3]	
Adult words (AWC)	228.5	(231.7)	118.5	[40.5; 377]	284.4	(253.7)	212	[69; 445]	– 55.8	(125.3)	– 30	[– 109.5; 23.2]
Child vocalizations (CV)	39.4	(39.6)	24	[8.8; 56.5]	38.5	(26.4)	37.5	[15.8; 58.2]	0.9	(25.6)	0	[– 8; 7.8]
Conversational turns (CT)	8.4	(7)	7	[3; 12]	22.9	(21.9)	17.5	[2.8; 38]	– 14.6	(19.2)	– 10	[– 28.2; 1]
Electronic media (TV)	31.1	(70)	0	[0; 5.8]	76.7	(121.8)	0	[0; 186.1]	– 45.6	(106.2)	0	[– 91.9; 0]

Mdn = Median, *Q1* = 1st quartile, *Q3* = 3rd quartile.

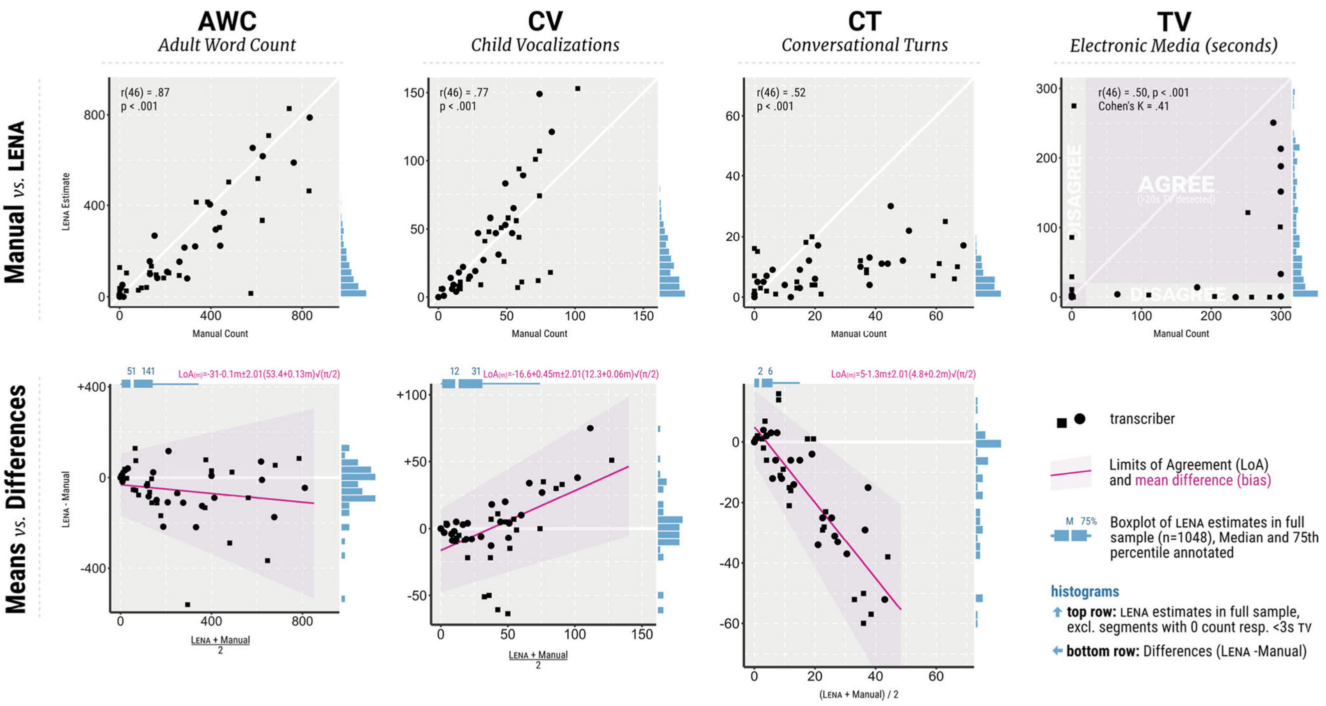

Fig. 1. Manual and LENA counts for $n = 48$ five-min segments. (Top row) Scatterplots of LENA estimates and manual counts, and histograms of LENA estimates in the full sample. (Bottom row) Bland–Altman plots of means against differences with limits of agreement (LoA), histograms of differences, and boxplots of LENA estimates in the full sample.

showed no significant difference between the CV counts from LENA ($Mdn = 24$) and the CV counts from the transcribers ($Mdn = 37.5$), $V = 485$, $p = .91$. We found a strong correlation between LENA's CV and the transcriber's utterance counts, $r(46) = .77$, $p < .001$.

The regression of the differences on the means revealed a significant proportional bias ($b_{1\,,\,\mathrm{diff}} = 0.45$, $t(46) = 4.42$, $p < .001$), with constant variance of the differences [$b_{1\,,\,\mathrm{res}} = 0.06$, $t(46) = 0.81$, $p = .42$)]. Hence, LoA were calculated as

$$\mathrm{LoA}(m) = -16.6 + 0.45m \pm 2.01 \times (12.3 + 0.06m)\sqrt{\pi/2}.$$

Again, we used this model to calculate the expected mean difference and LoA for samples with a varying number of child vocalizations. Interestingly, the model predicted the LENA counts to be lower than the manual counts for audio samples with $m \leq 36$ vocalizations, and higher otherwise. Specifically, the predicted mean difference for a sample with 36 vocalizations was −0.4, with LoA of ± 37 around that. According to the LENA estimates, 80% of the 5-min samples in the full sample had a CV count below this threshold. The median of LENA-estimated CV in the full sample was 12. For a sample with 12 vocalizations, the predicted mean difference was – 11, with LoA of ± 33 CV around that (i.e., – 44 to + 22; Fig. 1).

Conversational turns (CT)

On average, LENA counted fewer CTs than the transcribers, $M_{\mathrm{diff}} = -14.6$, $SD_{\mathrm{diff}} = 19.2$. A Wilcoxon signed rank indicated that the CT counts from LENA ($Mdn = 7$) were significantly lower than the manual CT counts ($Mdn = 17.5$), $V = 127$, $p < .001$. There was a moderate correlation between the LENA and manual CT counts, $r(46) = .52$, $p < .001$.

We found a significant proportional bias [$b_{1\,,\,\mathrm{diff}} = -1.3$, $t(46) = -11.13$, $p < .001$], indicating that the tendency of LENA to count fewer CTs than the transcribers increased with the number of CTs in the sample. The variance of the differences was also not constant [$b_{1\,,\,\mathrm{res}} = 0.19$, $t(45) = 2.85$, $p = .007$]. Thus, LoA were calculated as a function of the number of conversational turns in the 5-min sample m,—that is, as $\mathrm{LoA}(m) = 5.0 - 1.3m \pm 2.01 \times (4.8 + 0.2m)\sqrt{\pi/2}$.

As before, we used this model to calculate the expected mean difference for 5-min audio samples with a varying number of conversational turns. In the full sample, the median LENA-estimated CT was 2. For samples with two vocalizations, the model predicts the LENA counts to be 2.5 CTs higher than the manual counts, with LoA of ± 13 around that (i.e., – 10.5 to + 15.5). As the number of turns increases, the predicted mean difference becomes smaller and the LoA become wider. For samples that contain four or more turns, the LENA counts were predicted to be lower than the manual counts. At $m = 6$ (the 75th percentile of LENA CT counts in the full sample), the expected difference would be – 2.5 CT, with LoA of ± 14.5 (i.e., – 17 to + 12; Fig. 1).

Electronic media (TV)

On average, LENA counted less TV than the transcribers, $M_{\mathrm{diff}} = -45.6$ s, $SD_{\mathrm{diff}} = 106.2$. LENA detected TV mostly in samples in which the transcribers indicated uninterrupted TV—that is, close to 300 s. A Wilcoxon signed rank indicated that the difference between the TV counts from LENA ($Mdn = 0$) and the transcribers ($Mdn = 0$) was significant, $V = 67$, $p = .01$. We found a moderate correlation between the human and the LENA counts, $r(46) = .50$, $p < .001$.

The nonnormal distribution of the differences (Fig. 1) did not allow for calculation of LoA. Instead, we used Cohen's κ to assess how well the methods agreed on whether electronic media was or was not present (whereby "present" was defined as more than 20 s of audible electronic media). In 30 five-min samples, LENA and the transcribers found no TV; in seven samples, both did; for three samples, only LENA indicated TV; and for eight, only the transcribers did. This resulted in a Kappa of $\kappa = .41$.

Discussion

In the time it takes to transcribe a few minutes, LENA analyzes a day-long recording of a child's natural language environment. When large amounts of recording time need to be processed or when one is pressed for time, LENA could be a useful alternative to human transcription—provided that its measures are reliable. Whether that is the case was the main question of this study. Moreover, we wanted to assess LENA's reliability for Dutch. To this end, we investigated the correlation and agreement between LENA and manual transcription.

Correlation between LENA and manual transcription

Our first research question was how well the LENA counts for Dutch natural language recordings correlated with manual transcriptions, in particular as compared to other languages.

We found a strong correlation between manual and LENA AWC ($r = .87$). This is slightly lower than the $r = .92$ reported for English by Xu et al. (2009), but higher than the $r = .71$ and $r = .85$ reported by Oetting et al. (2009). Studies using other languages found lower correlations—namely, $r = .80$ for Spanish (Weisleder & Fernald, 2013), $r = .73$ for Chinese (Gilkerson et al., 2015), $r = .72$ for Korean (Pae et al., 2016), and $r = .64$ for French (Canault et al., 2016).

The correlation between human and LENA CV ($r = .77$) was slightly higher than the $r = .71$ reported for French by Canault et al. (2016).

The correlation for CT ($r = .52$) was lower than the values reported for Chinese, $r = .72$ (Gilkerson et al., 2015) and Korean, $r = .67$ (Pae et al., 2016). However, both of those studies excluded outliers to get significant correlations. Oetting et al. (2009), on the other hand, did not find a correlation between manual and LENA CT counts for English.

LENA's TV count did not correlate well with the manual transcriptions ($r = .50$) and often greatly underestimated the amount of TV. Correlations for TV have not been reported by other studies.

One explanation for the variation between studies might be the different languages that were used. The LENA algorithm has been trained on English. It might therefore perform better at estimating the number of adult words in languages that are similar to English—for example, in terms of prosody, phoneme inventory, or morpheme-to-word ratio. What exactly these similarities are is beyond the scope of this study to determine. However, the relative similarity between our results and those reported for English suggest that Dutch resembles English regarding the features that are relevant for the LENA algorithm.

Another potential explanation is methodological differences: For instance, some studies specifically selected samples with high speech activity (Gilkerson et al., 2015; Xu et al., 2009), whereas we and others (e.g., Canault et al., 2016; Weisleder & Fernald, 2013) used more diverse samples.

Studies also sample different environments: Around half of the recordings used by Pae et al. (2016) were of play and picture book reading in a hospital, whereas Oetting et al. (2009) used recordings they had made in a laboratory.

Moreover, Oetting et al. (2009) did not make the recordings with the LENA recorder, but merely played them back to it. Differences in microphone placement and the repeated recording and playback could have changed the signal in a way that made it more difficult to detect and differentiate speakers.

Participants' ages have differed, too. Canault et al. (2016), for example, included much younger children than we did (0–4 as opposed to 2–5 years). The correlations between LENA and human CT were particularly low for the younger age groups ($r = .49$ for 0–6 month olds and $r = .54$ for 7–12 month olds). This likely decreased the overall correlation relative to our study.

Agreement and biases between LENA and manual transcription

Our second research question was how well LENA agrees with manual transcription and whether there are any fixed or proportional biases between the two. We examined agreement and biases for AWC, CV, and CT by estimating mean differences and LoA (i.e., 95% prediction intervals for the differences between the two methods).

On average, LENA counted significantly fewer adult words. Some other studies have also reported lower LENA than manual AWC (Canault et al., 2016; Xu et al., 2009), whereas others found no significant difference (Gilkerson et al., 2015; Oetting et al., 2009).

The LoA for AWC were wide, and their width increased with the number of words in the sample. For a typical 5-min sample with 51 words, the estimated mean difference and LoA were $\mathrm{LoA}(51) = -36 \pm 152$ —that is, a relative difference of up to nearly 300%.

Although there was no significant mean difference between the human and LENA CV counts, the LoA were wide: -11 ± 33 CV for a 5-min sample with 12 vocalizations (the median of LENA CV in the full sample). Again, this was a difference of up to almost 300%. The tendency for human CV counts to be higher is in line with the results of Canault et al. (2016).

Lena's CT counts were significantly lower than the transcribers'. LoA for a 5-min sample with two turns were 2.5 ± 13 CT (i.e., up to over 600% difference). Significantly lower LENA CT counts were also reported by Oetting et al. (2009) and Berends (2015). However, Gilkerson et al. (2015) found no significant difference between CT counts from LENA and transcribers for Chinese.

The mean differences for CV and CT must be interpreted cautiously, because both counts were affected by significant proportional biases. Although LENA's CV counts were lower than the manual counts for most samples, they were higher for samples that contained many vocalizations (specifically, the upper 25% of the distribution of the LENA-estimated CV in the full sample). LENA's CT counts, on the other hand, were higher than the manual counts for samples with few turns, and lower for samples with many turns.

The agreement on whether or not there was TV in a sample was Cohen's $\kappa = .41$. This is lower than the $\kappa = .57$ found by Xu et al. (2009; as reported in Christakis et al., 2009). Gilkerson et al. (2015), on the other hand, reported agreement of just 2% on labeling TV.

Causes and consequences of disagreements and biases

If one wants to use LENA in a study, its measures should be comparable across families or points in time. Thus, our third research question was whether the comparability of LENA measures could be affected by LENA's biases and disagreements. To that end, it is also helpful to understand what would cause these differences.

Some disagreements certainly originate from random error—that is, independent of the true score, the subject, or any other differences between recordings. LENA uses only acoustical cues, and will therefore inevitably make mistakes. The transcribers—despite linguistic insight and context knowledge—were not infallible, either (interrater agreement: $\mathrm{ICC} = .93$, $.92$, and $.97$ for AWC, CV, and CT, respectively). Truly random errors are unlikely to create spurious results or jeopardize comparability across measurements (although they create noise that can make it more difficult to detect true differences).

Systematic biases, on the other hand, can cause serious problems. Proportional biases, in particular, will make changes of equal magnitude appear smaller or larger, depending on their starting point. As a consequence, LENA could misrepresent differences between subjects or changes within subjects. We found significant proportional biases for LENA's CV and CT counts.

One explanation for such biases is the different implementations of the counts. We did not mimic the LENA algorithm, but counted words, turns, and vocalizations as we would without it. Most notably, LENA operationalizes child language production (CV) as breath groups, whereas the transcribers counted child utterances—that is, they also considered semantic boundaries. The manual count of child utterances increased less steeply than the LENA CV. As a consequence, differences in CVs between more and less talkative children or within-subject changes could be exaggerated.

Despite their disagreement with the number of child utterances, LENA's breath-based CVs might be valid indicators of language production. It is even conceivable that the two metrics would be closely aligned in younger children, who produce shorter utterances than those in our sample. Yet, it remains unclear how developmental changes would affect this correspondence. As it stands, the two implementations seem to measure different things, and should not be treated as equal.

Another cause of disagreements may be that the transcribers used context knowledge to fill in gaps or resolve ambiguity. For example, when marking TV, they might have included moments with no audible TV on the basis of their understanding of the situation—that is, that it is unlikely that the TV was turned off for just a few seconds. Note, however, that LENA also uses the surrounding audio to improve the detection of electronic media (see Xu, Yapanel, Gray, Gilkerson, et al., 2008).

Another source of disagreements and biases is LENA's conservative treatment of overlapping, faint, and noisy speech, which apparently caused it to discount many utterances that the transcribers counted. An extreme example of this was a 5-min sample in which LENA had counted almost no adult words, whereas the transcriber counted close to 600. This sample contained a conversation between adults in the car, during which the key child was in the back seat. Others have demonstrated the effect of reverberation and noise on LENA's reliability more systematically (Canault et al., 2016; Xu et al., 2009).

This susceptibility to noise could also cause a proportional bias in the CT count if LENA excludes too many of the noisy adult and child utterances that make up the conversations. Indeed, in our sample, increases in conversational turns caused much smaller increases in LENA's CT count. This proportional bias could deemphasize differences between families and make it difficult to accurately assess changes in parent–child interactions—for example, following an intervention.

If differences in the LENA counts can be mere artifacts of differences in environmental noise, this could compromise the comparability of LENA measures. Since noise is entangled with other predictors of child development this could also confound study results: For example, the homes of families with low socio-economic status (SES) are more crowded and noisy (Evans, 2004). SES, in turn, predicts language development (Hoff, 2003; Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010). Similarly, a larger family size likely also means more noise and overlapping speech in the home. At the same time, a larger number of siblings can have a negative impact on development (Downey, 2001; Macaulay & Ford, 2013). Thus, when comparing children with different socio-economic backgrounds or family sizes, developmental differences that might look like the effect of a poorer language environment could actually be the result of noisier home environments or a stronger dilution of parental resources (Downey, 2001).

Similar issues could arise from any difference between recordings that might affect LENA's performance. For example, the child's and mother's voices are more likely to be confused when their fundamental frequencies are similar (VanDam & Silbert, 2016), and LENA's detection of electronic sounds could be thrown off track by a particularly high-quality TV. Our sample contained recordings of different homes, speakers, and situations, and the wide LoA we found could be the results of many such factors. When designing and interpreting a LENA study, one should be cautious of such potential confounds.

Ultimately, it is not clear from our data how the disagreements and biases would accumulate in a regular full-day LENA recording: The transcribed audio samples were chosen to cover as wide a range of values as possible. This was meant to increase the sensitivity to biases. Yet, over an entire LENA recording, the variables appear to be mostly confined to a smaller range in which LENA's bias is relatively low. If that is the case, the proportional biases and poor agreement in the more extreme regions would have little effect. The distribution of the LENA estimates across all $n = 1{,}084$ five-min audio samples can be seen in the boxplots in Fig. 1. Of course, given the biases described above, these are somewhat biased approximations of the true distributions.

Nevertheless, over the hundreds of 5-min samples in a full-day recording, even small differences could accumulate into large absolute differences, and if their direction or magnitude were determined by the properties of the recorded situations, this could introduce confounds when comparing LENA measures between subjects or points in time.

Limitations

This study has a number of limitations. Most importantly, our sample was too small to conclusively map the limits within which LENA measurements are comparable. That is, we did not have enough data to assess how background noise, number of speakers, and reverberation influence LENA. Furthermore, our sample included only six children, and their age range (2–5 years) did not cover the entire range for which LENA has been developed (2–48 months).

We also used only two transcribers, and their interrater agreement was good but not perfect. Consequently, the error in the manual counts might be larger than is ideal. More and better-trained transcribers and a more fail-safe transcription procedure might provide better estimates of the true counts.

Another limitation is that our implementation of the counts partially differed from that of LENA. As we discussed above, this was intentional, because we wanted to count words, utterances, and turns in the same way we would without LENA. Nevertheless, this could be considered an unfair comparison, since our transcribers counted something other than what is counted by LENA (e.g., child utterances rather than child vocalizations).

Then again, some of the transcription rules we adopted from the LENA algorithm might be undesirable in a research study. For example, we treated onomatopoeia, neologisms, and family-specific word forms as regular adult words, and counted child utterances that consisted only of babbling. Although these can easily be excluded from manual counts, this is impossible with LENA. Thus, depending on what kinds of words, utterances, or turns one wants to count, the disagreements with LENA might become smaller or larger.

Conclusions

We found evidence for good correlations between LENA's AWC, CV, and CT counts and their manual counterparts. We also found good average agreement between the methods, at least within the range that seems most relevant in practice. However, the differences between the LENA and manual counts varied widely (as indicated by the LoA), and there were strong proportional biases for more extreme counts. This calls into question the comparability of the LENA measures across different subjects and circumstances, which is vital for most study designs. Researchers who want to replace manual transcription with LENA should therefore be cautious of differences between measurement situations that might affect LENA's accuracy. Some of these have been discussed above. Despite such concerns, we have no doubt that LENA and similar technologies will shape the future of language acquisition research, and we are excited to see the developments to come.

Author note This study was approved by the Medical Ethical Committee of KU Leuven. This study was supported by the FP7 People Programme (Marie Curie Actions), REA Grant Agreement FP7-607139 (iCARE). We thank Brenda Rodiers and Olivia Vandewalle (Master students in speech pathology and audiological sciences, KU Leuven) for their meticulous transcriptions and valuable contributions to the development of the transcription protocol. We also thank the families that provided the LENA recordings.

References

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, *32*, 307–317.

Ambrose, S. E., VanDam, M., & Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and Hearing*, *35*, 139–147. doi:https://doi.org/10.1097/AUD.0b013e3182a76768

Berends, C. (2015). The LENA system in parent–child interaction in Dutch preschool children with language delay (Doctoral thesis). Universiteit Utrecht.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *327*, 307–310. doi:https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135–160. doi:https://doi.org/10.1191/096228099673819272

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer [Computer program]. Retrieved from www.praat.org/

Buschmeier, H., & Włodarczak, M. (2013). TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Proceedings der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung* (pp. 152–157). Bielefeld, Germany.

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENA^{TM}) in European French. *Behavior Research Methods*, *48*, 1109–1124. doi:https://doi.org/10.3758/s13428-015-0634-8

Carstensen, B. (2010). Comparing clinical measurement methods: A practical guide. Chichester: Wiley.

Caskey, M., & Vohr, B. (2013). Assessing language and language environment of high-risk infants and children: a new approach. *Acta Paediatrica*, *102*, 451–461. doi:https://doi.org/10.1111/apa.12195

Christakis, D. A., Gilkerson, J., Richards, J. A., Zimmerman, F. J., Garrison, M. M., Xu, D.,… Yapanel, U. (2009). Audible television and decreased adult words, infant vocalizations, and conversational turns. *Archives of Pediatrics and Adolescent Medicine*, *163*, 554. doi:https://doi.org/10.1001/archpediatrics.2009.61

Downey, D. B. (2001). Number of siblings and intellectual development. *American Psychologist*, *56*, 497–504. doi:https://doi.org/10.1037/0003-066X.56.6-7.497

Evans, G. W. (2004). The environment of childhood poverty. *American Psychologist*, *59*, 77–92. doi:https://doi.org/10.1037/0003-066X.59.2.77

Ford, M., Baer, C. T., Xu, D., Yapanel, U., & Gray, S. (2008). The LENA^{TM} language environment analysis system: Audio specifications of the DLP-0121. Retrieved from http://lena.org/wp-content/uploads/2016/07/LTR-03-2_Audio_Specifications.pdf

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*, 141–151. doi:10.11613/BM.2015.015

Gilkerson, J., Coulter, K. K., & Richards, J. A. (2009). Transcriptional analyses of the LENA Natural Language Corpus (September 2008), 1–10.

Gilkerson, J., & Richards, J. A. (2009). The LENA natural language study (LENA Foundation Technical Report, September 2008), 1–26. Retrieved from www.lenafoundation.org/wp-content/uploads/2014/10/LTR-02-2_Natural_Language_Study.pdf

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F.,… Topping, K. (2015). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech Language and Hearing Research*, *58*, 445. doi:https://doi.org/10.1044/2015_JSLHR-L-14-0014

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*, 1368–1378. doi:https://doi.org/10.1111/1467-8624.00612

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*, 343–365. doi:https://doi.org/10.1016/j.cogpsych.2010.08.002

Johnson, K., Caskey, M., Rand, K., Tucker, R., & Vohr, B. (2014). Gender differences in adult–infant communication in the first months of life. *Pediatrics*, *134*, e1603–e1610. doi:https://doi.org/10.1542/peds.2013-4289

Ludbrook, J. (1997). Comparing methods of measurements. *Clinical and Experimental Pharmacology and Physiology*, *24*, 193–203. doi:https://doi.org/10.1111/j.1440-1681.1997.tb01807.x

Macaulay, C. E., & Ford, R. M. (2013). Family influences on the cognitive development of profoundly deaf children: Exploring the effects of socioeconomic status and siblings. *Journal of Deaf Studies and Deaf Education*, *18*, 545–562. doi:https://doi.org/10.1093/deafed/ent019

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk (3rd ed.). Mahwah: Erlbaum.

Oetting, J. B., Hartfield, L. R., & Pruitt, S. L. (2009). Exploring LENA as a tool for researchers and clinicians. *ASHA Leader*, *14*, 20. doi:https://doi.org/10.1044/leader.FTR3.14062009.20

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D.,… Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*, 13354–13359. doi:https://doi.org/10.1073/pnas.1003882107

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J. A., Ma, L., & Topping, K. (2016). Effects of feedback on parent-child language with infants and toddlers in Korea. *First Language*, *36*, 549–569. doi:https://doi.org/10.1177/0142723716649273

R Development Core Team. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Revelle, W. (2016). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois. Retrieved from http://cran.r-project.org/package=psych

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. doi:https://doi.org/10.1037/0033-2909.86.2.420

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE*, *8*, e80646. doi:https://doi.org/10.1371/journal.pone.0080646

Suskind, D., Leffel, K. R., Hernandez, M. W., Sapolich, S. G., Suskind, E., Kirkham, E., & Meehan, P. (2013). An exploratory study of "quantitative linguistic feedback": Effect of LENA feedback on adult language production. *Communication Disorders Quarterly*, *34*, 199–209. doi:https://doi.org/10.1177/1525740112473146

Thiemann-Bourque, K. S., Warren, S. F., Brady, N., Gilkerson, J., & Richards, J. A. (2014). Vocal interaction between children with

Down syndrome and their parents. *American Journal of Speech-Language Pathology*, *23*, 474. doi:https://doi.org/10.1044/2014_AJSLP-12-0010

VanDam, M., & Silbert, N. H. (2016). Fidelity of automatic speech processing for adult and child talker classifications. *PLoS ONE*, *11*, 1–13. doi:https://doi.org/10.1371/journal.pone.0160588

Wang, H. (2007). English as a lingua franca: Mutual intelligibility of Chinese, Dutch and American speakers of English (Doctoral thesis). Leiden University.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science*, *25*, 1314–1324. doi:https://doi.org/10.1177/0956797614531023

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152. doi:https://doi.org/10.1177/0956797613488145

Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated analysis of child phonetic production using naturalistic recordings. *Journal of Speech, Language, and Hearing Research*, *57*, 1638–1650. doi:https://doi.org/10.1044/2014_JSLHR-S-13-0037

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA^{TM} Language Environment Analysis System in young children's natural home environment* (LENA Technical Report LTR-05-2). Retrieved from http://lena.org/wp-content/uploads/2016/07/LTR-05-2_Reliability.pdf

Xu, D., Yapanel, U., Gray, S., & Baer, C. T. (2008b). *The LENA^{TM} language environment analysis system: The interpreted time segments (ITS) File* (LENA Technical Report LTR-04-2). Retrieved from http://lena.org/wp-content/uploads/2016/07/LTR-04-2_ITS_File.pdf

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

Zhang, Y., Xu, X., Jiang, F., Gilkerson, J., Xu, D., Richards, J. A.,… Topping, K. J. (2015). Effects of quantitative linguistic feedback to caregivers of young children: A pilot study in China. *Communication Disorders Quarterly*, *37*, 16–24. doi:https://doi.org/10.1177/1525740115575771