Reliability of the Language Environment Analysis (LENA) in French-English Bilingual Speech

Adriel John Orena[1, 3], Krista Byers-Heinlein[1, 2, 3, 4], Linda Polka[1, 3]

[1]School of Communication Sciences & Disorders, McGill University, Montréal, QC, Canada

[2]Department of Psychology, Concordia University, Montréal, QC, Canada

[3]Centre for Research in Brain, Language and Music, Montréal, QC, Canada

[4]Centre for Research in Human Development, Montréal, QC, Canada

Address for correspondence:

Adriel John Orena

School of Communication Sciences & Disorders

2001 McGill College Avenue, 8th floor

Montréal, Québec, H3A 1G1

E-mail: adriel.orena@mail.mcgill.ca

Telephone: +1 (514) 398 1210

Keywords: Bilingualism, Speech Input, Tool Assessment, LENA

Conflict of Interset: The authors have no conflict of interest to declare.

Funding statement: This research was funded by the Social Sciences and Humanities Research Council (410-2015-0385) to Polka, L, the ASA Raymond H. Stetson Scholarship in Phonetic and Speech Science to Orena, A.J., and support from the Concordia University Research Chairs program to Byers-Heinlein, K.

Abstract

Purpose: This study examined the utility of the Language ENvironment Analysis (LENA) recording system for investigating the language input to bilingual infants.

Method: Twenty-one French-English bilingual families with a 10-month-old infant participated in this study. Using the LENA recording system, each family contributed three full days of recordings within a one-month period. A portion of these recordings (945 minutes) were manually transcribed, and the word counts from these transcriptions were compared against the LENA-generated adult word counts.

Results: Data analyses reveal that the LENA algorithms were reliable in counting words in both Canadian English and Canadian French, even when both languages are present in the same recording. While the LENA system tended to underestimate the amount of speech in the recordings, there was a strong correlation between the LENA-generated and human transcribed adult word counts for each language. Importantly, this relationship holds when accounting for different-gendered and different-accented speech.

Conclusions: The LENA recording system is a reliable tool for estimating word counts, even for bilingual input. Special considerations and limitations for using the LENA recording system in a bilingual population are discussed. These results open up possibilities for investigating caregiver talk to bilingual infants in more detail.

1. Introduction

A key factor in shaping early language development is the nature of early speech to infants. Indeed, many studies have established that the quantity and quality of caregiver speech has a strong influence on various subdomains of language development, including vocabulary growth (e.g., Hart & Risley, 1995; Ramírez-Esparza, García-Sierra, & Kuhl, 2014). Thus, evaluating and improving techniques for analyzing caregiver speech is important for many professionals, including researchers, clinicians and educators. Before the advent of modern recording systems, much of the research on caregiver talk was conducted with observation sessions and diary reports, which were limited by construct validity, parental bias, and human resources. A recent technological development from the Language ENvironment Analysis (LENA) foundation has attempted to mitigate these concerns, and has thus been the focus of many recent research and intervention programs focused on caregiver talk to young children – initially with monolingual English children, but more recently with other various populations, including monolingual learners of other languages, children with identified disorders, and children from a range of socioeconomic backgrounds (see Wang et al., 2017 for a review). Here, we focused on another understudied but populous group: bilingual infants. Indeed, more and more children around the world are learning two languages from birth, but many aspects of bilingual language acquisition remain poorly understood. The purpose of this study is to assess the utility of the LENA recording system for examining caregiver talk to bilingual-learning infants.

The LENA recording system is a language measurement and analysis tool that assesses the auditory environment of a young child. The child-friendly hardware consists of a small 3 oz. wearable digital recorder that fits in the front pocket of a t-shirt or vest. The portable digital

language processor (DLP) allows for the recording to take place in the infant's natural environment (i.e., at home), and without the need for bulky or visually salient equipment or extra personnel to be present. The DLP can record up to 16 hours of audio, which is then uploaded to software that automatically analyzes and segments the audio file. The LENA recording system generates reports about a child's language environment based on patented algorithms in the software. The report quantifies several aspects of speech in the recording, including the number of words spoken by adults, the number of vocalizations by the child, the number of conversational turns between the child and adult speakers, and the amount of noise and TV sounds in the background. The recording system was originally designed to support clinical and educational programs by providing quick feedback to caregivers, with the ultimate goal of increasing talk between caregivers and their children. Indeed, preliminary reports show that involvement in a program that uses the LENA recording system has positive effects for a child's language development (e.g., Suskind et al., 2013; Gilkerson, Richards & Topping, 2017). It has also been used by researchers to examine caregiver language input and children vocalizations in a range of ages and population type (Wang et al., 2017).

In the short time span since its inception, the use of the LENA recording system in research has both confirmed important data and revealed new findings about language acquisition. The central focus of recent research is quantifying speech heard by different populations (e.g., preterm infants: Caskey, Stephens, Tucker, & Vohr, 2014; children with Down's Syndrome: Thiemann-Bourque, Warren, Brady, Gilkerson, & Richards, 2014; children with Autism Spectrum Disorder: Oller et al., 2010; older adults: Li, Vikani, Harris, & Lin, 2014), and exploring how changes in the amount of caregiver speech might affect different aspects of language development or speech perception. Several studies have confirmed the seminal findings

of Hart and Risley (1995), that showed a tight link between the quantity of caregiver speech and monolingual infant's vocabulary development (e.g., Caskey et al., 2014; Gilkerson & Richards, 2008). Other studies have discovered a link between aspects of caregiver speech and speech processing efficiency (Weisleder & Fernald, 2013), and with children's brain responses to speech sounds (Garcia-Sierra et al., 2011; Romeo et al., 2018). Relevant to the current study, a small handful of studies have found the same type of input effects in young Spanish-English bilingual children (Speech processing efficiency: Marchman, Martínez, Hurtado, Grüter, & Fernald, 2016; Brain responses to speech sounds: Garcia-Sierra, Ramírez-Esparza & Kuhl, 2016).

An important consideration for the utility of the LENA recording system is the accuracy of its generated reports. The LENA recording system was originally developed for use in English environments, as the speech analyzer is based on American English recordings (Xu, Yapanel & Gray, 2009). Indeed, the LENA-generated adult word counts for English input are strongly correlated with word counts generated by a human transcriber ($r = .92$, $p < .01$; Xu et al., 2009). Given the many children in the world who grow up in other types of language backgrounds, it is of great interest to understand how well the algorithms, particularly for word counts, perform in non-English language environments. Indeed, word forms are indexed differently across different languages, and it is possible that the algorithms developed for counting words in English may not be generalizable to other languages. Nevertheless, several studies have shown that LENA algoirthms can reliably count words in other languages, including European French (Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2015), Spanish (Weisleder & Fernald, 2013), Mandarin (Gilkerson et al., 2015), Korean (Pae et al., 2016), and Dutch (Busch, Sangen, Vanpoucke, & Wieringen, 2017). These studies found that the LENA-generated adult word counts were not necessarily accurate; they tended to underestimate the amount of word counts in

the recordings. Nevertheless, all of these studies report that the estimated adult word counts for these other languages are sufficiently reliable, with correlation coefficients between .64 (European French) and .84 (Spanish). These findings suggest that the underestimation of LENA-generated adult word counts is consistent enough across participants to be used as a tool to compare the quantity of input across different infants.

Even though the LENA estimates for adult word counts have been validated for different languages, generalizing these results to a bilingual environment is not necessarily straightforward. First, it is difficult to assess whether the LENA-generated estimates are more reliable for one language over the other in bilingual input, especially given that previous validation studies have only examined the reliability of different languages separately, and these studies varied in their validation methodologies, including sample size and transcription period selection (see Ganek & Eriks-Brophy, 2017 for a short review), as well as data analysis (Busch et al., 2017). As an example, consider the two languages represented in this study: English and French. Although these two languages share many cognates, they also differ in many linguistic properties (phonology, prosodic forms, speech rate; e.g., Abercrombie, 1967; Pellegrino, Coupé, & Marsico, 2011). There have been efforts to validate the LENA-generated estimates of word count for both English and French (Xu et al. 2009; Canault et al., 2015); however, it is possible that the LENA algorithms are reliable at counting words within each of these languages, but are not comparable across the two languages.

Secondly, there are aspects of a bilingual input that might weaken the reliability of the LENA estimates. Indeed, the bilingual environment is not the sum of two monolingual environments (Grosjean, 1989), and validating the LENA system for two languages separately does not necessarily mean it would be reliable for the two languages in a bilingual environment.

While the acoustics of two languages in a bilingual context can be as distinct as when measured separately (Danielson, Seidl, Onishi, Alamian & Cristia, 2014), there can still be varying degrees of accented speech and code-mixing in bilingual environments (e.g. Byers-Heinlein, 2012). In the original LENA technical report, Xu and colleagues (2009) acknowledged that the performance of the LENA algorithms can be affected by many sources of variability, including speaker variations (i.e., speaking style, speaking rate, speaker accent, pitch). Given that these speaker variations are often present in bilingual environments, it is important to examine how the LENA system deals with these challenges.

In sum, the goal of this study was to validate LENA's algorithms in a bilingual environment. As part of a larger project investigating input effects on language skills in bilingual infants from Montréal, Canada, we have collected LENA recordings of the auditory environment of 10-month-old infants hearing French and English. Here, we assessed the reliability of the LENA-generated adult word counts across different languages (English vs. French), genders (female vs. male), and accents (accented vs. non-accented speech). While there is evidence of age-related changes in infant-directed speech (Kalashinkova & Burnham, 2018), previous validation studies have not found a difference in the reliability of the adult word count estimates in the environments of children at different ages (e.g., Busch et al., 2017; Canault et al., 2015; Xu et al., 2012). Thus, we believe that examining the reliability of the LENA-generated estimates for adult word count at one age point (i.e., in 10-month-olds) woud be sufficient for generalizing across different ages of the child. In the discussion, we report our assessment of the utility and limitations of LENA for investigating input in bilingual homes.

2. Methods

2.1. Participants

The participants were 21 families with a 10-month-old infant (13 males, 8 females), who were recruited as part of a large-scale study examining the language input to bilingual-learning infants. All of the families lived in the Greater Montréal area in Québec, Canada, where over half of the population have knowledge of both French and English (57.44%; Statistics Canada, 2016). Parents reported no auditory or developmental neurocognitive disorders for their infants.

Family Information

During their first visit to the lab, the mean age of the infants was 9 months 29 days (range = 9 m 15 d – 10 m 14 d). All families consisted of one father (age range 27 to 46 years, M = 36.24) and one mother (age range = 30 to 41 years, M = 34.85). Ten infants had one older sibling, and two infants had two older siblings; the other nine infants were first-born single children. Mothers had an average of 17.9 years of education (range = 11 – 23), while fathers had an average of 17.1 years of education (range = 14 – 22). Our sample of families was from a mid to high socioeconomic background, with an average Hollingshead score of 52.2 (range = 31 – 66 out of a possible score of 66). Most of the infants were cared for at home full time by one or both of the caregivers (n=16), while a few spent significant time at home with a nanny (n=2) or were enrolled in full-time daycare (n=3).

During the family's visit to the laboratory, we conducted a language background interview with the family to estimate their child's exposure to each language (Byers-Heinlein et al., in press). Parents did a month-to-month breakdown of their child's exposure to each of their languages, followed by a lifetime estimate of proportion exposure to each language. Then, we calculated the average of these two values. Based on these parent-report measures, there were 12

infants in a French-dominant language environment (range = 58 - 75% exposure to French), and 9 infants in an English-dominant language environment (range = 57 - 76% exposure to English). Four families reported a small amount of exposure to a third language: Arabic, Kannada, Portuguese, and Spanish (1 - 5% exposure to the third language).

All caregivers reported having knowledge of both French and English, but the families diverged in terms of the strategies they used for speaking to their children. Some families reported that both caregivers spoke both English and French to their infant (n=9), others reported that one caregiver spoke both English and French and the other caregiver spoke only one language to their infant (n=8), and some reported using the one-parent, one-language strategy for their child (n=4). Overall, the participants represented a range of bilingual family language configurations.

2.2. Procedure

The first appointment took place in the laboratory. During this first session, we explained the purpose and the procedure to the families, gave them the materials for the study, and conducted the language background interview with one (n = 17) or both (n = 4) of the caregivers. Families were given three digital language processors (DLPs) and three vests designed to hold the DLPs. With these materials, families were asked to complete three full days of recordings at home – two on weekdays, and one on the weekend. Most families were able to follow this schedule, except for one family that recorded one weekday and two weekend days, and one family that recorded three weekdays.

At the end of each day of recording, parents were also asked to complete a daily activity diary to indicate the day's general activities. When the families finished doing the recordings, we scheduled another appointment – either at their home (n=19) or at the laboratory (n=2) – to pick

up the materials, complete more language questionnaires, and conduct a final interview. In all but two cases, both caregivers were present for the final appointment. The majority of families completed the session within two weeks (M = 12.71 days, SD = 5.59, range = 5 – 26 days).

2.3. Data analysis overview

The project provided us with 1008 hours of audio recordings (21 families X 3 days X 16 hours). Given the human resources and time constraints, we implemented a two-stage protocol for selecting periods of the recordings from each family to transcribe. These protocols were modeled after methods used in previous studies (Marchman et al., 2017; Ramírez-Esparza et al., 2014). Accordingly, our protocol was designed to select periods of the recordings that contained different languages (French vs. English vs. Mixed), different social interaction types (speech to infant vs. speech between caregivers), and different degrees of speech density (high amount of speech vs. low amount of speech). In doing so, we aimed to transcribe and analyze periods that were representative of different scenarios in the whole recording.

Data coding and recording selection

First, we constructed a coding scheme to extract information from the recording about who was speaking and in what language. The LENA output provides an estimate of total speech heard by the infant, but it does not differentiate between what languages are being spoken and to whom the speech is being directed. The coding scheme was inspired by the Infant Social Environment Coding of Sound Inventory (SECSI; Ramírez-Esparza et al., 2014). We divided the recordings into 30-second chunks via Audacity software. Relevant to this study, trained research assistants listened to each chunk and tagged each chunk for speaker context (i.e., who is speaking and to whom: *Mother, Father, Sibling, Infant, Other*), and language context (i.e., what language was being spoken: *English, French, Mixed, Unknown*). For the language context, if more than

one language was being spoken during the 30-second chunk, research assistants coded it as "Mixed"; and if it was not clear what language was being spoken, we tagged it as "Unknown". To streamline this process, we only coded chunks that contained speech per LENA measures; in other words, chunks that had zero word counts per LENA measures were not coded. Further, based on pilot analyses, we found that coding half of the recordings was sufficient for capturing the language breakdown of the recording. Thus, we decided to code *every other* 30-second chunk that contained speech. Seven research assistants completed this part of the project. Each research assistant was a simultaneously and highly proficient bilingual of French and English. Each completed a training file before coding recordings to be submitted for analyses. We assessed coder reliability for these training files, and found high reliability for tagging the speaker ($M_{reliability}$= 94.2 %; Range = 91.8 % - 96.4%), and the language ($M_{reliability}$= 92.4 %; Range = 88.1 % - 96.1%). As part of their training, each research assistant was assigned to jointly code with one other research assistant for one coding session in order to maintain consistency across coders.

After this time-intensive process, we organized and aggregated the data into five-minute samples, so that we could identify the total amount of speech in each of these samples (per LENA measures), divided by social interaction type (per human coders), and in what languages (per human coders). For each participant, we identified nine samples of 5-minute recordings that would be representative of different degrees of speech density (approximately the top 1st percentile and 50th percentile of adult word counts), directed to different interlocutors (infant, non-infant), and directed in different languages (English, French, and mixed). In total, we selected 189 five-minute samples (945 minutes of recordings) from 21 families to be transcribed.

Data transcription

Three research assistants completed the transcriptions after undergoing extensive training. One research assistant did a second pass of all transcriptions to ensure accuracy and consistency across files. The five-minute samples were transcribed in CLAN (Computerized Language ANalysis) software using modified CHAT transcription format. The LENA-generated file was processed through CLAN, which provides a framework for transcription. The file was segmented by the LENA algorithms into utterance-length segments, and it contained general information about who was speaking in each segment (Male adult speaker vs. Female adult speaker vs. Infant, as identified by LENA algorithms), and how many words LENA estimated were present for that specific segment. The research assistants orthographically transcribed each segment. They also identified the language of each segment, and counted the number words uttered in each language. A word was counted as such if it contained at least one syllable. We borrowed the rules set forth by Canault and colleagues (2015) to count words in both English and French. Specifically, free morphemes (e.g., *English*: the, a, an; *French*: le, la, les), prepositions (e.g., *English*: in, to, on; *French*: à, de, par), and pronouns (e.g., *English*: I, he, she; *French*: je, il, elle) were counted as one word. In addition, the elided forms were counted as part of the word to which it was attached (e.g., *English*: I'll, can't; *French*: aujourd'hui, l'chien).

3. Results

3.1. Descriptive statistics

There was considerable variability in the amount of talk among participating families. Per LENA's algorithms, our group of infants heard an average of 15,651 words from adults per day (*SD* = 8,190), with a wide range of 1,644 to 49,022 words in a single day.

We transcribed 189 5-minute samples (21 participants X 9 samples), for a total of 945 minutes of transcribed samples. Based on LENA algorithms, there were 50,419 LENA-generated

segments within these 5-minute samples, of which 31.3% were female adult speech, 14.7% were male adult speech, 14.1% were infant speech, 6.3% were sibling speech, 2.0% were electronic sounds, and the remaining 31.65% were either other noise sounds or silence. We also coded the activities during these samples, and found that the primary activity in these selected samples varied, including play time (n = 56), conversations between adults (n = 41), bath or dressing time (n = 35), meal time (n = 22), story time (n = 27), media time (n = 2), and other housekeeping activities (n = 6).

Table 1

Descriptive statistics of LENA-generated and transcribed adult word counts (AWC).

Sample type	Number of samples	LENA-generated AWC	Transcribed AWC	Wilcoxon test
		Mean (SD)	Mean (SD)	*V*
All regions	189	339.51 *(179.22)*	423.34 *(210.76)*	3411 **
English	176	155.94 *(168.44)*	192.30 *(189.65)*	3746 **
Female	162	124.64 *(147.75)*	153.78 *(165.37)*	3133.5 **
Male	102	71.12 *(116.01)*	85.33 *(121.15)*	1494.5 **
French	179	160.74 *(154.12)*	198.17 *(177.01)*	3040 **
Female	167	107.38 *(131.48)*	133.08 *(150.73)*	2982 **
Male	112	96.68 *(119.31)*	115.88 *(120.40)*	1185 **
Mixed	106	18.46 *(40.23)*	19.50 *(36.76)*	2328.5
Female	77	20.08 *(46.32)*	20.07 *(42.28)*	1544.5
Male	46	8.92 *(11.72)*	11.09 *(11.02)*	350.5 **

Note. The Wilcoxon signed rank test was used to compare the two measures (LENA-generated AWC and Transcribed AWC). ** asterisks represent significance at the $p < .05$ level.

3.2. Reliability of LENA for Adult Word Counts

Similar to previous studies, we observed that the adult word count (AWC) measure by the LENA software was an underestimate of the transcribed word count (see Table 1). On average, the words counted by LENA were 85.3% of those counted by transcribers (*SD* =

39.3%). Following previous studies, we used a non-parametric test (Wilcoxon signed rank test) to statistically compare these two measures. The Wilcoxon signed rank test is the appropriate statistical test to conduct when the distribution of values is not normally distributed. Results showed that these two measures are significantly different when comparing all samples [V = 3411, $p < .001$]. Table 1 shows that these measures are also significantly different within specific types of recorded input based on language (*English* and *French*) and LENA-tagged sex of the voice (*Female* and *Male*). These measures were not significant within samples that contained *Mixed* speech [V = 2328.5, p = .11]. Note, however, that there were not many mixed utterances within samples that contained language mixing, with a mean LENA-generated AWC of 18.46 (SD = 40.23) and a mean Transcribed AWC of 19.50 (SD = 36.76).

By-language analysis

To examine the relationship between LENA-generated AWC and transcribed AWC, we conducted a repeated measures correlation. This test was chosen to examine the correlation between two variables, while accounting for intra-individual differences across participants. This analysis revealed that these two measures are strongly correlated [$r_{rm}(167) = .77$, $p < .001$] (see Figure 1a), suggesting that the LENA-generated AWC is a reliable measure of the actual AWC in the recordings.

We then assessed the relationship between automated and manual word count measures when separating the data by language. For each sample, we identified the segments that our researchers tagged as either English-only, French-only, and language-mixed (containing both English and French words). For these segments, we compared the LENA-generated versus the transcribed AWC. Note that some samples did not include any English-only segments (n=10), French-only segments (n=13), language-mixed segments (n=83), and were excluded from the

respective analyses. The repeated measures correlation analyses reveal a significant relationship between the two measures for English-only segments [$r_{rm}(154) = .90$, $p < .001$], French-only segments [$r_{rm}(157) = .94$, $p < .001$], and language-mixed segments [$r_{rm}(84) = .97$, $p < .001$] (see Figures 1b, 1c and 1d), suggesting that the LENA algorithms were equally reliable in all language contexts.

We next examined whether this relationship between LENA-generated and transcribed word counts is stronger for one language over the other. To do so, we fit a linear mixed model. This analysis is appropriate for our dataset as we had repeated observations per participant, an unbalanced number of observations per language, and multiple random effects. The model had transcribed word count as the dependent variable, and LENA-generated word count and language as fixed effects, and participant and sample type as random effects. Prior to running the model, the language variable was rescaled and centered around zero. As expected, there was a significant main effect of LENA-generated word counts [$\beta = 1.08$, $t = 43.33$, $p < 0.001$], indicating a strong relationship between transcribed and LENA-generated word counts in both languages. There was no main effect of language [$\beta = -11.41$, $t = -.98$, $p = .33$], indicating no significant difference in transcribed word counts between the English and French segments. Critically, there was no significant interaction between the LENA-generated word count and language [$\beta = .07$, $t = 1.26$, $p = .21$], suggesting that the relationship between the transcribed and LENA-generated word counts was consistent across language contexts.

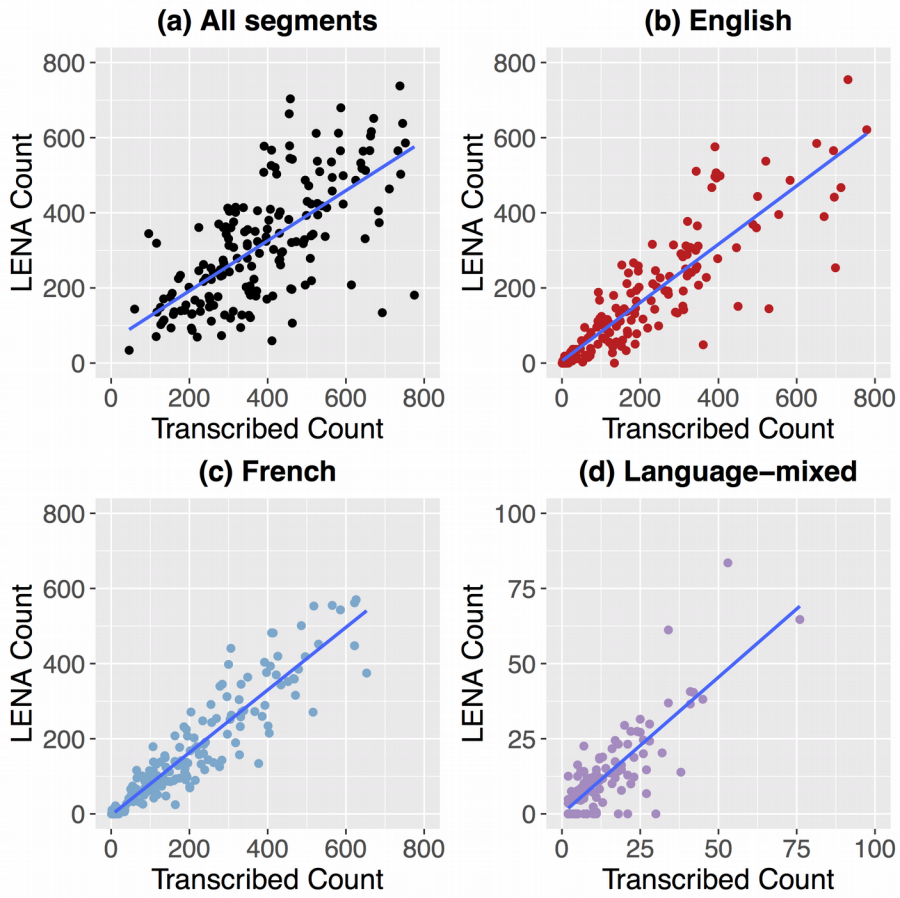

Figure 1. The relationship between LENA-generated and human-transcribed adult word counts for (a) all segments, (b) English-only segments, (c) French-only segments, and (d) Language-mixed segments. Each dot represents the word counts for one 5-minute sample.

By-gender analysis

We also examined whether the algorithms were similarly reliable for estimating word counts from female and male voices. To do so, we identified the segments that came from female and male voices, and then calculated an average for both the LENA-generated and transcribed AWC for each sample for each subject. Some samples did not include female (n=6) or male (n=55) voices, and were excluded from the respective analyses. Correlation analyses reveal a

strong positive correlation between LENA-generated and transcribed AWC for both female voices [$r_{rm}(161) = .85$, $p < .001$] and male voices [$r_{rm}(112) = .90$, $p < .001$].

To examine this relationship more closely, we conducted another linear mixed model for this dataset, with the transcribed word count as the dependent variable, and LENA-generated word count and gender as fixed effects, and participant and sample type as random effects. Again, there was a significant main effect of LENA-generated word count [$\beta = .93$, $t = 28.39$, $p < 0.001$]. There was no main effect of gender [$\beta = -19.23$, $t = -1.22$, $p = .22$], nor an interaction between LENA-generated word count and gender [$\beta = -.07$, $t = -1.06$, $p = .29$], indicating that the relationship between the transcribed and LENA-generated word counts was consistent across the different-gendered voices.

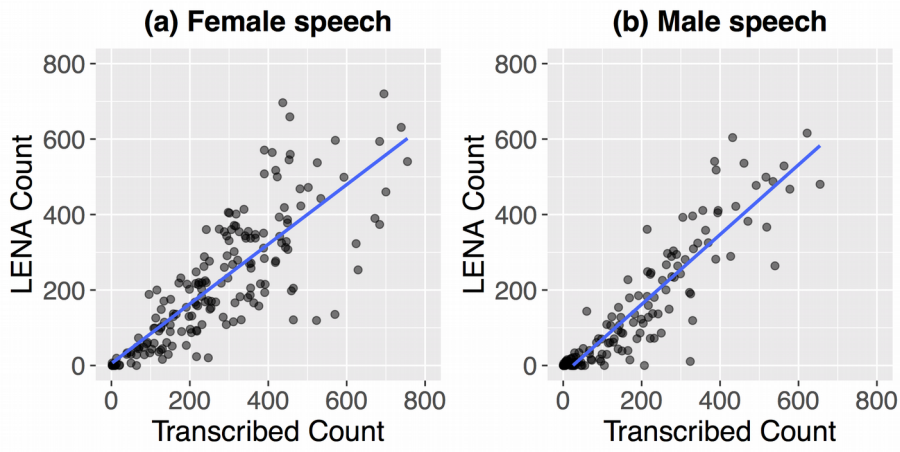

Figure 2. The relationship between LENA-generated and human-transcribed adult word counts for (a) segments with Female speech, and (b) segments with Male speech. Each dot represents the word counts for one 5-minute sample.

By-accent analysis

Finally, we examined whether the algorithms would be similarly reliable for estimating word counts from non-accented and accented speech. During the interview period, we asked

parents whether they felt that they had an accent in either of their languages. For mothers, seven reported that they did not have an accent in either language, eight reported that they had an accent in English, and six reported that they had an accent in French. For fathers, three reported that they had no accent in either of their languages, fourteen reported that they had an accent in English, and four reported that they had an accent in French. Correlational analyses reveal a strong, significant relationship between LENA-generated AWC and transcribed AWC for both accented speech [$r_{rm}(217) = .88$, $p < .001$] and non-accented speech [$r_{rm}(375) = .86$, $p < .001$].

To examine whether the reliability of the LENA-generated AWC changes with accented speech, we conducted a linear mixed model, with the transcribed word count as the dependent variable, and LENA-generated word count and accent type as fixed effects, and participant and sample type as random effects. As expected, there was a significant main effect of LENA-generated word count [$\beta = .96$, $t = 41.19$, $p < .001$]. However, there was no main effect of accent type [$\beta = 1.97$, $t = .35$, $p = .73$], nor an interaction between these two factors [$\beta = -.02$, $t = -.84$, $p = .40$]. These data indicate that the LENA algorithms fare well for counting words even in accented speech.

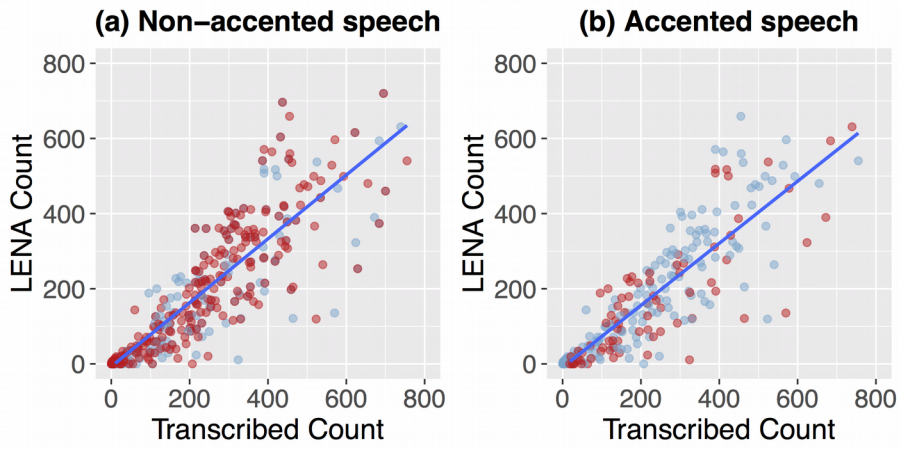

Figure 3. The relationship between LENA-generated and human-transcribed adult word counts for (a) segments with accented speech, and (b) segments with non-accented speech. Each dot represents the word counts for one 5-minute sample, with red dots representing English speech and blue dots representing French speech.

4. Discussion

In this study, we examined the utility of the LENA system for quantifying language input to bilingual infants. Specifically, we assessed the accuracy of LENA algorithms for performing adult word counts in bilingual speech. The languages spoken by the families were Canadian English and Canadian French; these two languages differ in many phonological, prosodic, and acoustic features, providing a test of whether LENA the algorithms, which were developed for American English, are applicable to languages with different features. Further, as is typical in bilingual settings, infants heard speech that contained both language mixing and accented speakers, which are features that may affect the reliability of LENA. Despite the potential challenges of measuring input in a bilingual context, we found high correlations between the LENA-generated and the human-transcribed adult word counts for both English and French – rs = .90 and .94, respectively – correlation coefficients which are similar to those found by other researchers who examined LENA performance in different monolingual language environments.

When we conducted separate analyses for different gendered voices and for accented speech, we also found a significant correlation between the LENA-generated and transcribed adult word counts. To our knowledge, this is the first study to directly compare the accuracy of the LENA system in a dual-language environment, and indeed, our results show that the algorithms in the LENA system adapt well to a bilingual speech stream.

There are still some broad challenges in the utlity of the LENA system that would be present even in monolingual contexts. In their original technical report, Xu and colleagues (2012) noted some systematic sources of error for the performance of the LENA algorithm, including the presence of background noise, overlapping speech, clothing of the child, and so forth. Indeed, in our own exploratory analysis, we found that the overall correlations between LENA-generated and transcribed adult word counts were stronger for infants without siblings (r_{rm} = .88, p < .001) versus infants with siblings (r_{rm} = .69, p < .001). Developing tools for minimizing these sources of errors would be beneficial for accurate and efficient speech analyses.

There are also some challenges of using the LENA system in bilingual contexts. First, LENA currently has no way of automatically classifying utterances according to different languages. Thus, while it can provide information on input in aggregate, it cannot do so by language. As described in the methods, it was necessary for us to manually tag the language in the recordings, which was quite time-intensive and likely impractical for clinical purposes. For practical reasons, we recommend conducting language interviews with caregivers to estimate a child's proportional exposure to each language, and then combining this language exposure proportion with the LENA-generated measures to obtain a volumetric estimate per language. For example, if a child heard 10,000 words per day (as measured by LENA), and their parent

reported that the child is exposed to French 60% of the time and English 40% of the time, then we would estimate that the child hears 6,000 words in French and 4,000 words in English. Indeed, preliminary results suggest that bilingual parents are sufficiently reliable at estimating their infant's language environment at home (Marchman, Martínez, Hurtado, Grüter, & Fernald, 2016; Orena, Byers-Heinlein & Polka, submitted), particularly when interviewers follow procedures to elicit high-quality information (Byers-Heinlein et al., in press).

Secondly, the linguistic landscape of bilingual homes is very heterogeneous, and it may be challenging to capture a representative sample of this experience, even when using the LENA system. For example, bilingual children tend to receive their dual-language input from more than one speaker, which necessitates obtaining consent from more individuals to gather a representative sample. At the minimum, this study required the involvement of both parents, especially since some families reported using a version of the one-parent, one-language strategy. In addition, bilingual children tend to receive their dual-language input from other sources as well, including grandparents, daycares, or community centers. Some families reported that they changed some of their activities during the day because they did not want to bring the recorders outdoors in public or because they did not want to have to involve their friends or family, which may have resulted in differences from the amount of each language the child might have heard on a more typical day. Further, issues related to obtaining consent prevented us from recruiting some families who had their child in full-day daycares. If researchers or clinicians are interested in exploring the bilingual input as a whole, they may need to observe the child in a wider variety of contexts.

Nonetheless, our results indicate that using the LENA recording system in bilingual homes does yield a valid representation of the bilingual language landscape. At the end of the

experiment, parents were asked a series of questions about their experience with the LENA recording device. When asked whether they felt that they changed the amount of speech they spoke to their child, only 14 out of 40 caregivers said "yes"[1]. Of these caregivers, ten said that they spoke to their child more often. For example, one parent noted that "*Once in a while, … if it was quiet, I'd be like "oh yeah, we're recording!" and I guess I should say something, so I would talk a tiny bit more.*" Four other participants said that they spoke to the child less often, indicating that they were a bit more cautious about discussions within the family. For example, one parent cited the *Hawthorne effect*, noting that *"there's something about… having a microphone that makes you more conscious about what you're doing".* Nevertheless, these parents reported that, while they acted differently at the beginning of the recordings, they quickly resumed their daily activities at home.

While we encouraged parents to act as natural as possible during the recordings, some parents did note that they changed the language proportion that they spoke to their child (n = 7). For example, one parent said, *"It [the study] almost just serves to remind me, like, I need to speak more English, and I'm hoping it, like, continues from now on"*, and one other parent said, *"I probably changed by putting more English to it [the recording]… It's more of a realizing that, oh, maybe we don't do it [speak English] as much."* This suggests that, in addition to increasing the quantity of language input at home, the tool could also be used as a way to change the language proportion at home to be closer to parents' goals. Indeed, several studies have shown that differential patterns of language exposure affect various speech and language outcomes, including speech processing skills (Hurtado et al., 2013) and vocabulary development (Thordardottir, 2011). Thus, it may be beneficial for parents to know the raw amount of input

[1] Note that two caregivers were not present for the post-study interview; thus, we only report interview data for 40 out of 42 caregivers in this study.

that their child hears in each language. Future studies should examine the potential utility of this tool for improving bilingual language outcomes in children.

While our study provides strong support for using the LENA system in a bilingual context, this study also had some limitations. First, our sample included families from medium to high socio-economic (SES) backgrounds, most of whom provided care to their children at home (note that many families in Montréal enrol their children in daycare around 12 months of age, due to government parental leave policies). Thus, our sample is not representative of all bilingual families in Montréal or globally. While the current study is focused on validating the LENA estimates of AWC (and not examining the quality of family language use), it is possible that there were less background noise or overlapping speech in our recoding samples (which are potential sources of error). Nevertheless, one previous study of Spanish-speaking families from primarily low-SES backgrounds also reported high reliability of the LENA adult word count algorithm (Weisleider & Fernald, 2013), suggesting that LENA would be valid in bilingual contexts across the SES spectrum.

Second, our study focused only on the accuracy of the LENA-generated word counts for Canadian English and Canadian French. There are several ways that future studies can expand on our findings to further test the flexibility of the LENA algorithms across different metrics and bilingual contexts. For example, future work could investigate the accuracy of LENA-generated frequencies of child vocalizations and turn-taking in bilingual settings. Prior work with non-English languages would suggest high accuracy of child vocalizations and turn-taking estimates in a bilingual context (e.g., Ganek et al., 2018; Gilkerson et al., 2015), but future research is needed to confirm these predictions. It is also important to examine whether the LENA

algorithms are also reliable with other language pairs that differ in more linguistic properties than English and French (e.g., a pair of tonal and non-tonal languages).

In sum, the LENA recording system offers researchers and practitioners a means to investigate a child's language input at home. Our study shows that the use of the system can be extended to French-English bilingual families. From a research perspective, this tool can answer more detailed questions about how specific parameters of the bilingual input affect language skills, which can inform theories of language development (Odean, Nazareth, & Pruden, 2015). For example, the LENA system provides an estimate of the absolute amount of input that a child hears, which has been shown to be predictive of a bilingual child's speech and language skills (Marchman et al., 2016; Garcia-Sierra et al., 2016). From a practical perspective, this tool can be used for improving caregiver talk to bilingual infants. Indeed, many children grow up learning two languages, so it is important to develop tools that can be used in both monolingual and bilingual environments. Certainly, the system has been shown to be effective in increasing caregiver talk, even in different cultures (Benítez-Barrera, Angley, & Tharpe, 2018; Pae et al., 2016). On both ends, the LENA recording system opens up possibilities for investigating caregiver talk to children and improving their child's development.

Acknowledgements

Our research would not be possible without the support of all the families who opened up their homes to us and took part in this study. We thank our research team, particularly Hicks, A., Higgins, F., and Kerr, S. for coordinating the recruitment of participants, Dang Guay, J., Deegan, M., Fabbro, J., Martel, S., Raftopoulos, A., Wang, E., and Xu, K. for coding the data, and Srouji, J., Lei, V. and Custo Blanch, M. for transcribing the data. We also thank the Centre for Research on Brain, Language and Music for providing us with the LENA equipment.

5. References

Abercrombie, D. (1967) *Elements of general phonetics*. Edinburgh: Edinburgh University Press

Benítez-Barrera, C. R., Angley, G. P., & Tharpe, A. M. (2018). Remote Microphone System Use at Home: Impact on Caregiver Talk. *Journal of Speech, Language & Hearing Research*, *61*(2), 399–409. https://doi.org/10.1044/2017_JSLHR-H-17-0168

Busch, T., Sangen, A., Vanpoucke, F., & Wieringen, A. van. (2017). Correlation and agreement between Language ENvironment Analysis (lena™) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*, 1–12. https://doi.org/10.3758/s13428-017-0960-0

Byers-Heinlein, K. (2012). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism*, *16*(1), 32–48. https://doi.org/10.1017/s1366728912000120

Byers-Heinlein, K., Schott, E., Gonzales-Barrero, A. M., Brouillard, M., Dubé, D., Jardak, A., … Tamayo, M. P. (in press). MAPLE: A Multilingual Approach to Parent Language Estimates. *Bilingualism: Language and Cognition.*

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2015). Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods*. https://doi.org/10.3758/s13428-015-0634-8

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2014). Adult Talk in the NICU With Preterm Infants and Developmental Outcomes. *Pediatrics*, *133*(3), e578–e584. https://doi.org/10.1542/peds.2013-0104

Danielson, D. K., Seidl, A., Onishi, K. H., Alamian, G., & Cristia, A. (2014). The acoustic

properties of bilingual infant-directed speech. *The Journal of the Acoustical Society of America, 135,* EL95. https://doi.org/10.1121/1.4862881

Ganek, H. V, & Eriks-Brophy, A. (2017). A Concise Protocol for the Validation of Language ENvironment Analysis (LENA) Conversational Turn Counts in Vietnamese. *Communication Disorders Quarterly*. https://doi.org/10.1177/1525740117705094

Garcia-Sierra, A., Ramírez-Esparza, N., & Kuhl, P. K. (2016). Relationships between quantity of language input and brain responses in bilingual and monolingual infants. *International Journal of Psychophysiology*. https://doi.org/10.1016/j.ijpsycho.2016.10.004

Gilkerson, J., & Richards, J. A. (2008). The LENA foundation natural language study (Technical Report LTR-02-2). Retrieved from: http://www.lenafoundation.org/TechReport.aspx/Natural_Language_Study/LTR-02-2

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*(2), 445–452. https://doi.org/10.1044/2015_JSLHR-L-14-0014

Grüter, T., Hurtado, N., Marchman, V. A., & Fernald, A. (2014). Language exposure and online processing efficiency in bilingual development. *Input and Experience in Bilingual Development*, *13*, 15. https://doi.org/10.1075/tilar.13.02gru

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H Brookes Publishing.

Kalashnikova, M. & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar properties but different functions. *Journal of Child Language, 45*(5), 1035-1053. https://doi.org/10.1017/S0305000917000629

Li, L., Vikani, A. R., Harris, G. C., & Lin, F. R. (2014). Feasibility study to quantify the auditory and social environment of older adults using a digital language processor. *Otology and Neurotology*, *35*(8), 1301–1305. https://doi.org/10.1097/MAO.0000000000000489

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & Fernald, A. (2016). Caregiver talk to young Spanish-English bilinguals: comparing direct observation and parent-report measures of dual-language exposure. *Developmental Science*. https://doi.org/10.1111/desc.12425

Odean, R., Nazareth, A. & Pruden, S. M. (2015). Novel methodology to examine cognitive and experiential factors in language development: combining eye-tracking and LENA technology. *Frontiers in Psychology, 6*(1266). https://doi/org/10.3389/fpsyg.2015.01266

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*(30), 13354–13359. https://doi.org/10.1073/pnas.1003882107

Orena, A. J., Byers-Heinlein, K., & Polka, L. (submitted). What do bilingual infants actually hear? Evaluating measures of caregiver speech to 10-month-olds.

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*(6), 549–569. https://doi.org/10.1177/0142723716649273

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language, 87*(3), 539-558. https://doi.org/10.2307/23011654

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech

development. *Developmental Science*, *17*(6), 880–891. https://doi.org/10.1111/desc.12172

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2016). The Impact of Early Social Interactions on Later Language Development in Spanish-English Bilingual Infants. *Child Development*. https://doi.org/10.1111/cdev.12648

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science, 29*(5), 700-710. https://doi.org/10.1177/0956797617742725

Statistics Canada (2016). Focus on Geography Series, 2016 Census. *Statistics Canada Catalogue no. 98-404-X2016001*. Ottawa, Ontario.

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*(3), 474–485. https://doi.org/10.1044/2014_AJSLP-12-0010

Wang, Y., Hartman, M., Aziz, N. A. A., Arora, S., Shi, L., & Tunison, E. (2017). A Systematic Review of the Use of LENA Technology. *American Annals of the Deaf*, *162*(3), 295–311. https://doi.org/10.1353/aad.2017.0028

Weisleder, A., & Fernald, A. (2013). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Xu, D., Yapanel, U., & Gray, S. (2009). Reliability of the LENATM Language Environment Analysis System in young childrens natural home environment. *Behavior Research Methods* 48(3), 1109–1124. https://doi.org/10.3758/s13428-015-0634-8