Comparing different measures of bilingual input derived from naturalistic daylong recordings

Yufang Ruan*[1,2], Adriel John Orena[3], Linda Polka[1,2]

[1] School of Communication Sciences and Disorders, Faculty of Medicine and Health Sciences, McGill University, Montreal, Canada

[2] Centre for Research on Brain, Language and Music, Montreal, Canada

[3] Department of Psychology, University of British Columbia, British Columbia, Canada

Correspondence Author: Yufang Ruan, 2001 McGill College Avenue 8th floor, Montréal, Québec, H3A 1G1, Canada, yufang.ruan@mail.mcgill.ca.

How to cite this article: Ruan, Y., Orena, A. J., & Polka, L. (2023). Comparing Different Measures of Bilingual Input Derived From Naturalistic Daylong Recordings. *Journal of Speech, Language, and Hearing Research*, *66*(5), 1618-1630. https://doi.org/10.1044/2023_JSLHR-22-00180

Abstract

Purpose: Measuring language input, especially for infants growing up in bilingual environments, is challenging. Although the ways to measure input have expanded rapidly in recent years, there are many unresolved issues. In the current study, we compared different measurement units and sampling methods used to estimate bilingual input in naturalistic daylong recordings.

Method: We used the Language Environment Analysis (LENA) system to obtain and process naturalistic daylong recordings from 21 French-English bilingual families with an infant at 10 and 18 months of age. We examined global and context-specific input estimates and their relation with infant vocal activeness (i.e., volubility) when input was indexed by different units (Adult Word Counts, speech duration, 30-second segment counts) and using different sampling methods (every-other-segment, top-segment).

Results: Input measures indexed by different units were strongly and positively correlated with each other and yielded similar results regarding their relation with infant volubility. As for sampling methods, sampling every other 30-second segment was representative of the entire corpus. However, sampling the top segments with the densest input was less representative and yielded different results regarding their relation with infant volubility.

Conclusions: How well the input that a child receives throughout a day is portrayed by a selected sample and correlates with the child's vocal activeness depends on the choice of input units and sampling methods. Different input units appear to generate consistent results, while caution should be taken when choosing sampling methods.

Comparing different measures of bilingual input derived from naturalistic daylong recordings

Measuring language input, especially for infants growing up in bilingual environments, is challenging. Some researchers have used diaries and questionnaires completed by caregivers to estimate the proportion of each language in a child's input (e.g., Carbajal & Peperkamp, 2020; Place & Hoff, 2011, 2016). Other researchers have documented children's real-world input using audio- or video-recordings. In recent years, a growing number of researchers have adopted the Language Environment Analysis system (LENA, LENA Research Foundation, Boulder, CO) to obtain and process daylong audio recordings of language input in bilingual households (e.g., Orena et al., 2020; Marchman et al., 2017; Ramírez-Esparza et al., 2017a, 2017b; VanDam et al., 2016). The LENA system includes a recorder that children wear in a vest and algorithms that automatically process and estimate language input by adult word counts (AWCs) or speech duration. Researchers have tested LENA's accuracy in different languages and have found most input estimates to be fairly reliable (Cristia et al., 2020, 2021; Orena et al., 2019).

Although the ways to measure input have expanded rapidly, there are still many unresolved issues. First, various units have been used to measure bilingual input. Some researchers, using diary or recording methods, have divided time in a day into equal-sized segments and measured segment counts or durations where each language was used (e.g., Place & Hoff, 2011, 2016; Ramírez-Esparza et al., 2017a, 2017b). This method of *counting segments* is widely used for its simplicity and efficiency. However, bilingual caregivers do not always use the same language within a given segment, especially in a relatively longer segment (e.g., 30 minutes). Thus, in some cases, researchers have asked caregivers to estimate the proportion of time that each language was used within a segment (Carbajal & Peperkamp, 2020). This approach improves the accuracy of bilingual exposure estimation, but still overlooks the fact that caregivers might not continuously speak for the length of a segment. This limitation can be addressed by using more fine-grained units, such as *speech duration* or *AWCs* extracted from recordings via speech processing algorithms (e.g., Marchman et al., 2017; Ruan, Orena, & Polka, 2020; Ruan, Orena, Xu, et al., 2020). However, the accuracy of these algorithms is imperfect (Cristia et al., 2020, 2021; Lehet et al., 2021), and these fine-grained units are not always available (e.g., when using diaries and questionnaires). Therefore, it is important to examine whether using different units impacts input estimation. Until we do so, it is difficult to compare results across studies where input was indexed by different units.

Another unresolved issue is that algorithms are so far unable to automatically and reliably classify bilingual input into two languages; this task requires manual annotation which is a laborious endeavor. This challenge underscores the need for an effective and reliable sampling method which allows researchers and clinicians to achieve their goals by processing only a portion of their data. Orena and colleagues (2020) examined daylong recordings in the Montréal Bilingual Infants corpus and found that the proportion of bilingual input in each language estimated from a periodically-selected (e.g., *every-other-30s-segment sampling*, EOS) sample was well-correlated with parental estimations. Taking the bilingual input proportions estimated from an EOS sample as the gold standard, another study further showed that a smaller randomly-selected sample (7% of the total recording or 11% of speech clips) yielded an estimation close to the gold standard (Cychosz et al., 2021). This study suggests that a reliable estimation of the bilingual input distribution can be achieved by annotating a modest amount of recording. Others have tried to achieve the same goal using more selective sampling methods. For example, in one study, researchers focused on the most input-dense portions of the recordings by selecting 40 temporally-scattered segments *with the highest AWCs* each day (i.e., the top segments), from which they composed a sample of around 160 top segments across four days for each child (Ramírez-Esparza et al., 2017b). However, there are concerns around this *top sampling method*: the features of the input in the top segments may differ from what a child experiences throughout a day (Bergelson et al., 2019; Tamis-LeMonda et al., 2017). It is also unknown whether the distribution of a child's bilingual exposure remains the same in the top segments. Taken together, examining the representativeness of different samples is a crucial step towards identifying effective and reliable sampling methods.

Representativeness aside, researchers and clinicians may sample the top segments for different objectives. For example, children growing up in bilingual families are simultaneously exposed to two languages. Many bilingual children have more opportunities to receive input in one language (dominant) than the other (non-dominant). Thus, it is informative to control for the variance in the opportunity that a child has with each language when examining how language dominancy shapes the relation between the input and child vocal behaviours. Selecting the same number of segments and selecting the ones with the densest input for each language (e.g., top sampling of input in a specific language) is a way to obtain two comparable samples with the same duration and optimal density for language comparisons (Xu et al., 2019). The same sampling approach can also be used to compare input received in different social contexts. Because the implementation of this top sampling method requires annotating the input (i.e., in what language and what social context) prior to sampling, we call it post-annotation top sampling.

In our ongoing project, we are interested in using LENA recordings in the Montréal Bilingual Infants corpus to investigate the relation between infants' vocal activeness (i.e., volubility) and their French-English bilingual input in different social (overhearing, one-on-one) and language (dominant, non-dominant) contexts when infants were 10 and 18 months old. Following Orena et al. (2019), we plan to estimate input using the AWCs from a sample of every other 30-second segment containing adult speech across the entire corpus. Prior to that, the reliability of this method was examined in the current study. Specifically, we examined whether input measures indexed by AWC in the EOS sample were consistent with measures indexed using other units (speech duration, segment counts) and sampling methods (the entire corpus, simple top sampling, and post-annotation top sampling; see a summary of variables in Table 1). We also examined whether the distribution of different social and language input in the top segments resembled the distribution in the EOS sample. Lastly, we compared the relation between infant volubility and input when the input was estimated using different units and sampling methods.

Findings from this study can help aggregate results from studies using different input units and/or sampling methods as well as provide methodological guidance for future studies using daylong recordings. The findings can potentially be applied to all research regardless of participants' language background (monolingual, bilingual, etc.), while some will be particularly relevant to bilingualism research. For example, counting segments in each language is widely used to determine the bilingual input distribution, thus knowing the reliability of this method will have a wide implication. Moreover, the post-annotation sampling method derives comparable samples with equal duration and maximal density, which enable us to compare input in two languages while controlling for the inherent difference in quantity to some extent.

Methods

Participants

We analyzed data from the Montréal Bilingual Infants corpus (Orena et al., 2020). Twenty-one families participated when the infant was 10 months old (13 males, 8 females; Age *Mean* = 303 days, *Range* = 289 – 319 days) and 16 of them participated again when the child was 18 months old (10 males, 6 females; Age *Mean* = 576 days, *Range* = 551 – 635 days). All caregivers had knowledge of both French and English and most of them (27 out of 42) reported speaking both languages to their child. According to parental estimates, their child was exposed to each language for at least 20% of time. Four families reported a small amount of exposure to a third language (< 5%). At 10 months, 12 infants were raised in a French-dominant language environment and nine were English-dominant. At 18 months, eight were French-dominant and eight were English-dominant. Parents provided consent to participate and declared no auditory and neurocognitive disorders for their child.

Procedure and Measures

Measures used in this study are summarized in Table 1. Naturalistic audio recordings were collected using a LENA digital language processor (DLP). Infants wore the DLP in a vest for 16 hours per day. Three full-day recordings (2 weekdays and 1 weekend day) were made when infants were 10 months old. For 16 families, a fourth recording was completed on a weekend day when infants were 18 months old. In total, the families contributed 1,264 hours of audio recordings ([21 families at 10 months × 3 days × 16 hours] + [16 families at 18 months × 1 day × 16 hours]). Recordings were divided into 30-second segments. Estimates of child vocalizations and language input were derived for each segment using the LENA algorithms. The Child Vocalization Count (CVC) is the number of vocalizations produced by the key child. A child vocalization is defined as a speech/speech-like sound produced by the key infant that is preceded and followed by 300 milliseconds of silence or nonspeech. We summed CVCs across the entire corpus for each child at each age to index *infant volubility*. Language input was measured using different units and sampling methods. How each sampling method was conducted while using different units is described in Figure 1.

Input Units (AWC, Duration, Segment Count)

LENA algorithms estimate the number of words spoken near the key child (Adult Word Counts, AWC). Previous research showed that LENA algorithms were reliable at estimating AWCs in both English and French (Orena et al., 2019). Algorithms also estimate the duration of these words and derive Adult Female Speech Duration and Adult Male Speech Duration. For each infant, the sum of Adult Female and Male Speech Duration provided an approximation of speech duration (Duration). The Segment Count referred to the number of 30-second segments.

LENA Sample

The LENA sample consisted of all the recordings in the corpus for each age. There were 21 families × 3 days × 1920 segments per day = 120,960 segments in the 10-month LENA sample and 16 families × 1 day × 1920 segments per day = 30,720 segments in the 18-month LENA sample. One segment in the 10-month sample was excluded because of an evident technical error (AWCs > 3000 in a 30-second segment, accounting for less than 0.5% of total AWC). As the Segment Count was identical for all infants, we utilized AWC and Duration to measure input in the LENA sample.

Every-Other-Segment (EOS) Sample

As we were interested in caregivers' input, we first removed segments in the LENA sample that did not contain any adult speech. From the remaining segments containing adult speech, we selected *every other segment*. In total, 18,979 and 6,180 segments were included in the 10-month and 18-month EOS samples respectively.

Segments in the EOS sample were manually annotated. Trained English-French bilingual research assistants listened to each segment and coded for social contexts (i.e., how many speakers and listeners, who was speaking to whom) and language contexts (i.e., what language was being spoken). Seven research assistants completed this work after each of them successfully completed a training file. Inter-coder reliability in the training file was high (on average 94.2% agreement for speaker context and 92.4% agreement for language context, Orena et al., 2020). Speech in which one caregiver spoke directly to the infant was tagged as one-on-one input. Overheard input was tagged for speech spoken in the presence of the infant, but not directly addressing the infant. Speech in which two or more caregivers spoke directly to the infant was not included as another level of social contexts because this was rarely observed in our corpus. Input tagged as “English” or “French” were recoded as “dominant” or “non-dominant” with the dominance assigned according to the *parent-reported* relative exposure to each language for each child at each age. Mixed-language input was not included as another level of language contexts because it accounted for less than 10% of the total input on average at each age.

For each child at each age, we summed the total input in the EOS sample (global) and computed input measures by social contexts (one-on-one, overhearing) and language contexts (dominant, non-dominant). As the number of segments containing adult speech varied across infants, there was a considerable variation in the Segment Count in the EOS sample. Thus, we utilized all three units (Segment Count, AWC, and Duration) to measure input in the EOS sample.

Top150 Sample

Following the work of Ramírez-Esparza and colleagues (Ramírez-Esparza et al., 2017a, 2017b), we selected the top 50 segments with the highest AWCs each day across three days in the 10-month EOS sample for a total of 150 segments per child. For 18 months, despite having only one daylong recording, we sampled the top 150 segments with the highest AWCs for each child. This allowed us to examine whether the size of top samples relative to the original sample affects input estimation.

Because the Top150 sample was selected from the EOS sample, social and language context annotation was also accessible for the Top150 sample. Again, for each child at each age, we summed the total input in the Top150 sample (global) and indexed it in AWC and Duration (Segment Count was identical for all infants, n = 150). We also summed input by social and language contexts and used all three measurement units (Segment Count, AWC, and Duration) to index the input in each context.

Top40/20 Samples

Segments in the EOS sample were initially categorized by social (one-on-one or overhearing) and language (dominant or non-dominant) contexts according to the manual annotation. For the two social contexts, top 40 segments with the highest AWCs in each context were sampled for each child at each age. We chose 40 segments because most infants had at least 40 segments for each social context except one child at 18 months (one-on-one context analysis was based on 23 segments for this child). For the two language contexts, top 20 segments with the highest AWCs in each language were sampled for each child at each age. Again, we chose 20 segments because most infants had at least 20 segments for each language context with only a few exceptions (non-dominant language analysis was based on less than 20 segment for one child at 10 months (Segment Count = 7) and two children at 18 months (Segment Count = 4 and 9)). Together, at each age, we had one Top40 sample for each of the two social contexts (one-on-one and overhearing) and one Top20 sample for each of the two language contexts (dominant and non-dominant, see Figure 1). In total, we had four Top40/20 samples for 10 and 18 months respectively. By definition, Segment Count was identical for each context; thus, input was indexed only by AWC and Duration in the Top40/20 samples.

Statistical Analysis

Results and plots were generated using packages including languageR (Baayen & Shafaei-Bajestan, 2019) and ggplot2 (Wickham, 2016) in R (R Core Team, 2021). The data and code that support the findings of this study are available at https://osf.io/uqh35/.

To examine whether using different units provide similar input estimates, we correlated input measures in different units. Next, to investigate whether each sampling method generates a representative sample, we correlated input measures derived from a selected sample to the measures derived from its original sample. Spearman's correlations were used because the input distribution deviated from a normal distribution. Significance of these correlations was not tested because we were interested in the degree of these correlations (i.e., the magnitude).

Next, to examine whether the proportions of input in different social and language contexts remain the same in top segments as the ones observed throughout a day, we computed these proportions in the Top150 and EOS samples. We used AWCs in a specific context divided by the total AWCs in that sample. Then, for each context, we compared proportions estimated in the two samples using Wilcoxon signed-rank tests. All *p*-values were adjusted using method of Benjamini & Hochberg (1995).

Lastly, to test whether the relation between infant volubility and input changes depending on how input was estimated, we compared Spearman's correlations between infant volubility and language input when input was estimated using different units and sampling methods. We repeated the analysis in different social and language contexts at each age. The original *p*-values were reported because (1) The purpose of this set of analyses was not to test the hypothesis that infant volubility was related to language input, but to examine the consistency across input units and sampling methods; (2) We tried to mimic the reality where researchers and clinicians would only select one measure of input and there would not be any *p*-value adjustment at the level of input measurement.

Results

Does using different units and sampling methods provide similar estimations of language input from daylong recordings?

Spearman's correlations between different input measures are plotted in Figure 2, for global input (a) as well as input in each social (b & c) and language (d & e) context. Results for the 10-month dataset are plotted in the upper triangle and results for the 18-month dataset, in the bottom triangle. The Spearman's correlation coefficient between each pair of input measures is reported in each cell and the cell colour indicates the strength of the correlation, from weak (yellow) to strong (red). Conventionally, a value of .80 or greater indicates a good consistency across measures (Chiang et al., 2020, p. 98). A video-animated guide of Figure 2 is available in the *Supplementary Material*.

First, we compared across different units (AWC, Duration, and Segment Count). Within each sample, we correlated input indexed by different units. A stronger positive correlation indicated a higher consistency between two units. We expected the correlation between AWC and

Duration to be positive and strong, while the correlation of each with Segment Count to be less strong because Segment Count is a less fine-grained unit and less dependent on speech processing algorithms compared to AWC and Duration. As expected, we observed strong correlations across three units in all contexts, samples, and ages (shown in the cells close to the diagonal line in Figure 2 and in Part 1 of the animation). For example, in Figure 2a, the cell corresponding to the first column from the *left* (Column 1 or C1, LENA_AWC) and the second row from the *bottom* (Row 2 or R2, LENA_Dur) shows the correlation between the global input estimated in the entire 10-month corpus by AWC and speech duration, which is .99 (> .80) suggesting a good consistency between these two unites. Correlations involving Segment Count were relatively smaller, especially for overhearing context in Top150 samples (Figure 2b, [C4-5, R6] and [C6, R4-5]).

Next, we compared across different sampling methods (LENA, EOS, Top150, and Top40/20). The EOS sample was drawn from the LENA sample (i.e., the entire corpus) and the annotation of context-specific input was not available for the LENA sample, thus we examined the correlation between the global input estimated in the EOS and LENA samples (see Figure 2a, [C1-2, R3-5] and [C3-5, R1-2] and Part 2 of the animation). We expected these correlations to be positive and strong. Indeed, the correlation coefficients were beyond .80 (a few below this threshold involving Segment Count). These results indicated that the EOS sample selected by every-other-segment sampling was representative of the entire corpus.

The Top150 sample was drawn from the EOS sample, thus we examined the correlation between the input estimated in these two samples. We expected weaker correlations here because top sampling provides a narrow snapshot of the child's language exposure throughout a day (Bergelson et al., 2019). Indeed, our results showed that compared to the correlations between the EOS and LENA samples reported above, the correlations between the global input derived in the Top150 and EOS samples were slightly smaller (Figure 2a, [C3-5, R6-7] and [C6-7, R3-5], see also Part 3 of the animation). Although correlations were close to .80 for input in both language contexts (Figure 2d & e, [C1-3, R4-6] and [C4-6, R1-3]) and in the 18-month dataset (bottom triangles), correlations were below this threshold when input was indexed by Segment Count and for both types of social input in the 10-month dataset (Figure 2b & c, [C1-3, R4-6]). For instance, in 10-month dataset, the correlations between the overheard input estimated in the EOS sample and in the Top150 sample ranged from .34 to .75 (Figure 2b, [C1-3, R4-6]), evidently smaller than the threshold (.80), as well as smaller than their corresponding correlations in the 18-month dataset that ranged from .56 to .98 (Fugure2b, [C4-6, R1-3]). These results suggested that the most input-dense portions of the recordings might be less representative of the daylong recordings.

The Top40/20 samples were drawn from the EOS sample with the goal of equating the number and input density of the segments used to examine the variation across different types of input. We did not have a clear hypothesis for the correlation between each type of social and language input in Top40/20 samples and in the EOS sample as this analysis was exploratory. Except for the dominant language input in 18-month dataset and measures involving Segment Count, our results suggested a good representativeness of the Top40/20 samples by showing correlations close to .80 (Figure 2 b-e, [C1-3, R7-8] and [C7-8, R1-3], see also Part 4 of the animation). These results indicated that when we attempt to control for inherent differences in infants' opportunity to receive each type of input, we still observe a similar pattern of individual differences. Therefore, this post-annotation top sampling method can potentially provide a representative sample of specific types of input.

Do estimated proportions of input in different social and language contexts differ when different sampling methods are used?

Due to the discrepancies observed between context-specific input in the EOS and Top150 samples, we further compared the proportional estimates of social and language input across these two samples. Given that a previous study found different input patterns in peak-hour versus daylong samples (Bergelson et al., 2019), we expected the input proportions estimated in the Top150 sample to differ from the ones in the EOS sample.

As shown in Table 2, differences were observed between the two samples with median ranging from 1 to 7%, and they reached significance for overheard, one-on-one, and dominant language input in 10-month samples (original *p*s < .05). However, none of these *p*-values was significant after correcting for multiple comparisons. The results indicated that the input distribution across different social and language contexts in the top segments differed from the distribution observed throughout a day, but not substantially.

Does the relation between input and infant volubility change when the input is estimated using different input units or sampling methods?

We compared the correlation between infant volubility (derived from the entire corpus) and the input when the input was estimated by different units and sampling methods. When comparing across different units (AWC, Duration, Segment Count), similar correlations indicated consistency. We found that within the EOS sample, input-volubility correlations were generally same in direction and significance, and similar in magnitude across different units (see Table 3 columns 2-4). For example, the correlation between volubility and overheard input at 10 months was .49, .49, and .52 when input was indexed by Segment Count, AWC, and Duration respectively. These correlations were uniformly positive, significant at 0.05 level, and numerically close to each other. These results suggested that using different units to measure input led to similar conclusions regarding the relation between input and infant volubility.

When comparing across sampling methods (EOS, Top150), we viewed the input-volubility correlations where the input was estimated in the EOS sample as the gold standard. Thus, deviations from this gold standard suggested potential problems with the top sampling method. Based on the results from our previous research questions, we expected a discrepancy between the correlations for the EOS and Top150 samples. Indeed, compared to the EOS correlations (Table 3, columns 2-4), the Top150 correlations (columns 5-7) were consistently smaller and sometimes in the opposite direction (e.g., input in overhearing contexts at both ages and non-dominant language contexts at 18 months, all indexed by Segment Count). For example, compared to the correlation between volubility and overheard input estimated using AWC and EOS sampling at 10 months (*rho* = .49), the corresponding correlation was much smaller when input was estimated in the Top150 sample (*rho* = .28). Therefore, using a simple top sampling method to estimate input might lead to a different conclusion regarding the relation between input and infant volubility.

Additionally, we examined the correlation between infant volubility and the context-specific input estimated in Top40/20 samples. This post-annotation sampling method was used to achieve a different goal, namely, to assess whether the input-volubility correlation would change when comparable samples were used for each type of input. Therefore, if Top40/20 correlations deviates from the EOS gold standard, it would not suggest problem with this post-annotation sampling method but reveal that the quality of the input (language dominancy or social interaction) played a role independent from any inherent differences in the opportunity that a child has with a specific type of input. This comparison was exploratory hence we did not have a clear expectation. As shown in the last two columns in Table 3, the Top40/20 correlation coefficients were generally smaller than the corresponding EOS correlations (hence less likely to be significant). Meanwhile, the Top40/20 correlation coefficients were numerically closer to the EOS ones. For example, although the correlation between volubility and overheard input at 10 months when input was estimated using AWC and Top40 sampling was smaller and still not significant (*rho* = .39), this value was closer to the one observed for the EOS sampling (*rho* = .49). The pattern across two social contexts (overhearing versus one-on-one) as well as across two language contexts (dominant versus non-dominant) also generally reassembled the pattern observed in the EOS sample. These results advanced our findings from the first research question by showing that in addition to input's variation, its covariation with infant volubility also persisted in the Top40/20 samples.

Discussion

In summary, our analyses yielded the following findings: (1) Input measures indexed by different units (AWC, Duration, and Segment Count) were positively and strongly correlated with each other and generated similar results regarding their relation with infant volubility; (2) Input estimates derived using the every-other-segment sampling method were representative of input in the entire corpus; (3) Sampling the top segments with the densest input might derive a less representative sample for estimating input and its relation with infant volubility; and (4) Context-specific input's variation and its covariation with infant volubility persisted in segments selected by a post-annotation top sampling method.

Measures of language input using different units (AWC, Duration, and Segment Count) and their relation with infant volubility were highly consistent. Hence, we validated the method of using AWC to estimate the input in LENA daylong recording for our ongoing project.

Meanwhile, correlations involving Segment Count were slightly smaller and this deviation was amplified when we compared units across samples. For instance, in the 10-month dataset, the correlation between the Segment Count of overheard input in the Top150 sample and the AWC of the same type of input in the EOS sample was only .34, much smaller than .80 (Figure 2b, [C1, R6]). In addition, the only correlations that showed a negative relation between input and infant volubility, were based on input measures indexed by Segment Count (Table 3, column 5). These findings have important implications on how we assess bilingual exposure, given that counting segments or segment duration is a common practice in previous bilingualism research (e.g., Place & Hoff, 2011, 2016; Ramírez-Esparza et al., 2017a, 2017b). When counting segments, we lose information such as how verbally active the speaker is and whether the speaker consistently uses the same language for the entire segment. Some researchers have tried to address the latter by asking caregivers to estimate the time that each language was used within a segment (Carbajal & Peperkamp, 2020). In future studies, researchers could also estimate the time that caregivers are actively speaking within a segment to quantify the input more precisely, when fine-grained units like AWC and speech duration are not available. On the other hand, one of the advantages of using Segment Count is that counting segments relies less on speech processing algorithms, which spares it from concerns regarding the accuracy of these algorithms (Cristia et al., 2020, 2021; Lehet et al., 2021).

Our results also showed that sampling every-other-segment achieved a good representativeness of the entire corpus, which replicated our previous pilot study (Orena et al., 2019). Meanwhile, a sample of the top segments with the densest input was less representative of a child’s language exposure throughout a day, shown by the correlations between the Top150 and EOS samples (Figure 2a – e). In addition, the correlations between infant volubility and the input estimated in the Top150 sample diverged from the ones where the input was estimated in the EOS sample (Table 3). This deviation may arise for two reasons. One is biased sampling. The Top150 sample consisted of segments containing the highest AWCs, essentially the moments when caregivers were the most verbally active around the child (talking to the child or others). The distribution of different types of input might differ in top segments. Indeed, we observed that the proportion of overheard and one-on-one input differed between the Top150 and EOS samples in the 10-month dataset (although no longer significant after *p*-adjustment, Table 2). These differences corresponded to the weaker correlations observed in Figure 2 (b) and (c). Other aspects of the input in the top segments might also differ from infants' language experience throughout a typical day, as suggested by previous research where researchers found a denser usage of nouns in peak-hour recordings (Bergelson et al., 2019). These differences observed for the input in the top segments might help explain the deviant relation between infant volubility and input found when the input was estimated using the Top150 sampling method (Table 3). If it is true, periodic or random sampling which selects input without reference to input features, might yield a less biased sample (Cychosz et al., 2021; Orena et al., 2019).

The other possible reason is related to the size of the Top150 sample relative to its original sample (i.e., the EOS sample). Recall that albeit 10-month dataset (3-day recordings) being larger than 18-month dataset (1-day recording), we selected the same number of top segments (n = 150) for each child from these two datasets. Therefore, the Top150 sample accounted for a smaller proportion of the EOS sample for 10-month dataset (17%) than for the 18-month dataset (37%). This might explain why we observed relatively weaker correlations between input measures estimated in the Top150 and EOS sample in the 10-month dataset (Figure 2 upper triangles) than the 18-month dataset (Figure 2 bottom triangles). Therefore, sampling a fixed number of top segments might be disadvantageous for larger samples: For a given number of top segments, the larger the original sample is, the smaller proportion of segments are selected, and thus less likely to be representative. To tackle this problem with larger samples, it might be helpful to sample a fixed proportion, instead of a fixed number, of segments, so that the size of the selected sample would changes with the size of the original sample. Future studies should consider finding an optimal proportion for selecting a representative sample from daylong recordings with the least segments. For example, previous research suggested that around 7% randomly-selected segments from overall recordings was representative in terms of the proportion of bilingual exposure and child-directed speech observed in the EOS sample (Cychosz et al., 2021). However, the correlation between these estimations based on 7% of data and parental reports were still not optimal. We should also keep in mind that no matter how many segments are sampled from the recordings, samples only provide a snapshot of a child's everyday life hence carry some extent of sampling bias.

We also examined a post-annotation top sampling for a different purpose which was to assess whether the pattern of different types of social or language input and their relation with infant volubility would change when comparable samples with equal duration and maximal density (Top40/20 samples) were used for different types of input. Our results showed that when being provided with the same opportunity and in its optimal condition, each type of social and language input seems to preserve its variation and its covariation with infant volubility as observed throughout a day. These results also suggested that this post-annotation top sampling might be used for other purposes, such as more detailed annotations. For example, for researchers and clinicians who are interested in comparing child-caregiver interaction when caregivers speak one language or the other, they might consider to initially annotate language contexts (i.e., which language(s) was used) for every other segment containing adult speech, and then select a fix proportion of top segments from each language context to conduct detailed annotations on child-caregiver interactions.

Infant age might also contribute to the different patterns observed in the correlation between infant volubility and input received in different contexts. Although we presume that the differences can be primarily attributed to the fact that language input and infant volubility increase and the relation between them changes as infants grow, we cannot rule out potential impacts from our methodological choices. For example, the recordings were collected on a weekend day at 18 months while they were collected on three days including both weekdays and weekends at 10 months. Many factors including children's daily routine and primary caregiver(s) may vary across weekday and weekend. Additionally, when children were at 18 months, some of them went to a daycare but language input at the daycare was not captured in our recordings. Activities at settings outside home may alter children's language exposure (Larson et al., 2020; Soderstrom et al., 2018), which might potentially alter our results.

There are limitations to the current study. First, although previous reviews and evaluations have suggested the LENA-derived measures, especially Child Vocalization Counts (CVCs, indexed infant volubility) and AWCs, to be reasonably accurate (Cristia et al., 2020, 2021), there are still some gaps between the accuracy of manual annotation and automatic processing algorithms (Lehet et al., 2021). Second, our sample size is relatively small ($N = 21$ and 16 for 10- and 18-month respectively) due to the laborious work involved in manual annotation, but the corpus consists of 1,264 hours of daylong recordings. Third, the families contributed to the corpus lived in a French-English bilingual community in which both languages have high social status. This characteristic may restrict the generalization of findings to bilingual communities where language status is uneven. For example, we suspect that the top sampling methods may yield less-representative results in bilingual contexts where languages have different levels of social status and/or involve more complex patterns of language use.

There are several directions for future studies. First, future studies should try to replicate the findings of this study using a larger sample, in different bilingual contexts, and recordings made outside of the home. Second, when Ramírez-Esparza and colleague composed their sample using the simple top sampling method, the authors made the effort to ensure selected segments were 3-minute apart (Ramírez-Esparza et al., 2017a, 2017b). Whether this effort would improve the representativeness of top sampling remains to be tested. Third, although manually coding adult speech in every other segment reduced the work needed to code the full corpus by half, it was still laborious. Future studies could investigate the reliability of other periodic, but less dense sampling methods, such as sampling 1 minute every hour (Scaff et al., 2022).

In conclusion, while the methods to estimate children's language input have been expanding rapidly in recent years, it is important to know that our research conclusions are not built on methodological biases. Our results suggested a high consistency across different units (AWC, speech duration, segment count). However, caution should be taken when choosing sampling methods. While sampling every other 30-second segment might generate a unbiased sample, there is more work needed to be done to improve the representativeness of top sampling methods. That said, top sampling methods can still be used for different research purposes. Taken together, findings from this study highlight the need for our field to direct more attention to the exact measures used to estimate language input and to be thoughtful when selecting sampling methods.

Acknowledgement

We thank families who have contributed to the Montréal Bilingual Infant corpus, and research assistants at McGill Infant Speech Perception Lab for coordinating the research (Hicks, A., Higgins, F., Kerr, S., and Zhou, G.), coding the data (Dang Guay, J., Deegan, M., Fabbro, J., Martel, S., Raftopoulos, A., Wang, E., and Xu, K.), transcribing the data (Srouji, J., Lei, V. and Custo Blanch, M.). We thank Dr. José A. Correa for his expert statistical consultancy and our anonymous reviewers for their instrumental comments. This project is supported by the China Scholarship Council-McGill Joint Fellowship (201706040050) and Fonds de recherche du Québec – Société et culture Doctoral Research Scholarship (2022-B2Z-303687) to YR, and by grants from the Social Sciences and Humanities Research Council of Canada (410-2015-0385) and from the Centre for Research on Brain, Language and Music to LP.

Ethics Approval Statement

We received ethics approval from the Institutional Review Board at McGill University (IRB # A05-B20-16A).

Data Availability Statement

The data and code that support the findings of this study are available at https://osf.io/uqh35/.

Supplemental Material: https://doi.org/10.23641/asha.22335688

References

Baayen, R. H., & Shafaei-Bajestan, E. (2019). *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics* (1.5.0). https://cran.rstudio.com/web/packages/languageR/index.html

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. https://doi.org/10.1111/desc.12715

Carbajal, M. J., & Peperkamp, S. (2020). Dual language input and the impact of language separation on early lexical development. *Infancy*, *25*(1), 22–45. https://doi.org/10.1111/infa.12315

Chiang, I.-C. A., Jhangiani, R. S., & Price, P. C. (2020). *Research Methods in Psychology – 2nd Canadian Edition*. BCcampus. https://opentextbc.ca/researchmethods/

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis System Segmentation and Metrics: A Systematic Review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

Cychosz, M., Villanueva, A., & Weisleder, A. (2021). Efficient Estimation of Children's Language Exposure in Two Bilingual Communities. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3843–3866. https://doi.org/10.1044/2021_JSLHR-20-00755

Larson, A. L., Barrett, T. S., & McConnell, S. R. (2020). Exploring Early Childhood Language Environments: A Comparison of Language Use, Exposure, and Interactions in the Home and Childcare Settings. *Language, Speech, and Hearing Services in Schools*, *51*(3), 706–719. https://doi.org/10.1044/2019_LSHSS-19-00066

Lehet, M., Arjmandi, M. K., Houston, D., & Dilley, L. (2021). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system. *Behavior Research Methods*, *53*(1), 113–138. https://doi.org/10.3758/s13428-020-01419-y

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: Comparing direct observation and parent-report measures of dual-language exposure. *Developmental Science*, *20*(1), e12425. https://doi.org/10.1111/desc.12425

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2019). Reliability of the Language Environment Analysis Recording System in Analyzing French–English Bilingual Speech. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2491–2500. https://doi.org/10.1044/2019_JSLHR-L-18-0342

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2020). What do bilingual infants actually hear? Evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental Science*, *23*(2), e12901. https://doi.org/10.1111/desc.12901

Place, S., & Hoff, E. (2011). Properties of Dual Language Exposure That Influence 2-Year-Olds' Bilingual Proficiency: Dual Language Exposure and Bilingual Proficiency. *Child Development*, *82*(6), 1834–1849. https://doi.org/10.1111/j.1467-8624.2011.01660.x

Place, S., & Hoff, E. (2016). Effects and noneffects of input in bilingual environments on dual language skills in 2 ½-year-olds. *Bilingualism: Language and Cognition*, *19*(5), 1023–1041. https://doi.org/10.1017/S1366728915000322

R Core Team. (2021). *R: A language and environment for statistical computing.* (4.1.2). R Foundation for Statistical Computing. https://www.R-project.org/

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017a). Look Who's Talking NOW! Parentese Speech, Social Context, and Language Development Across Time. *Frontiers in Psychology*, *8*, 1008. https://doi.org/10.3389/fpsyg.2017.01008

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017b). The Impact of Early Social Interactions on Later Language Development in Spanish-English Bilingual Infants. *Child Development*, *88*(4), 1216–1234. https://doi.org/10.1111/cdev.12648

Ruan, Y., Orena, A. J., & Polka, L. (2020, July). *The Relation of Language Input and Infant Volubility in English-French Bilingual Families*. International Congress of Infant Studies 2020, Virtual. https://www.researchgate.net/publication/342715654_The_Relation_of_Language_Input_and_Infant_Volubility_in_English-French_Bilingual_Families

Ruan, Y., Orena, A. J., Xu, K., & Polka, L. (2020). Language input and volubility in French-English bilingual infants. *The Journal of the Acoustical Society of America*, *148*, 2501. https://doi.org/10.1121/1.5146940

Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2022). Characterizing children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. PsyArXiv. https://doi.org/10.31234/osf.io/mt6nz.

Soderstrom, M., Grauer, E., Dufault, B., & McDivitt, K. (2018). Influences of number of adults and adult: Child ratios on the quantity of adult language input across childcare settings. *First Language*, *38*(6), 563–581. https://doi.org/10.1177/0142723718785013

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science*, *20*(6), e12456. https://onlinelibrary.wiley.com/doi/10.1111/desc.12456

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. *SEMINARS IN SPEECH AND LANGUAGE*, *37*(2), 128–141. https://doi.org/10.1055/s-0036-1580745

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag. https://ggplot2.tidyverse.org.

Xu, K., Orena, A. J., Ruan, Y., & Polka, L. (2019). The effects of parental interaction on French-English bilingual infants' vocalization and turn-taking rates. *Journal of the Acoustical Society of America*, *145*, 1765. https://doi.org/10.1121/1.5101466

Tables

Table 1 *Variables and descriptions.*

Variables	**Descriptions**
Infant Language Development	
Infant volubility	LENA-derived child vocalization counts (CVC) in the entire corpus.
Language Input	
Measuring Units	
AWC	The LENA-derived estimate of the number of words spoken near the child.
Duration	The sum of LENA-derived Adult Female Speech Duration and Adult Male Speech Duration.
Segment Count	The number of 30-second segments.
Sampling Methods	
Every-other-segment sampling	A periodic sampling method selecting every other 30-second segment containing adult speech.
Top sampling	A sampling method selecting a certain number of segments with the highest AWCs. Two top sampling methods were examined in this study: a simple top sampling (see Top150) and a post-annotation top sampling (see Top40/20).
Samples	
LENA	The entire Montréal Bilingual Infants corpus, consisted of 1,264 hours of audio recordings.
EOS	The sample selected by every-other-segment (EOS) sampling method, consisted of every other 30-second segment containing adult speech from the entire corpus. Every segment was annotated for speaker(s), listeners(s), and language usage.
Top150	The sample selected from the EOS sample by the simple top sampling method, consisted of top 150 segments with the highest AWCs.
Top40/20	Samples selected from the EOS sample by the post-annotation top sampling method, consisted of top 40 segments with the highest AWCs in a specific social context (overhearing, one-on-one), or top 20 segments with the highest AWCs in a specific language context (dominant, non-dominant).
Social and Language Contexts	
Global	All input in the sample.
Overhearing	Caregivers spoke in the presence of the infant but not exclusively addressing the infant.
One-on-one	One caregiver (mother, father, nanny, older sibling, and other) talked to the infant.
Dominant language	Parent-reported language (French or English) that the infant has more exposure to at each age.
Non-dominant language	The language other than the dominant language (English or French).

Table 2 *Comparison of proportions of language input (indexed by AWC) in different social and language contexts across Every-other-segment (EOS) and Top150 samples* [1].

	EOS		Top150		Difference[1]		Wilcoxon V
	Median	Interquartile Range	Median	Interquartile Range	Median	Interquartile Range	
Social Contexts							
10M: Overhearing	73%	68 – 76%	81%	70 – 87%	6%	4 – 10%	43#
10M: One-on-one	25%	20 – 32%	19%	11 – 28%	7%	3 – 9%	183#
18M: Overhearing[2]	66%	44 – 82%	68%	44 – 82%	2%	0.8 – 4%	57
18M: One-on-one	34%	18 – 56%	32%	18 – 56%	-[2]	-	-
Language Contexts							
10M: Dominant	51%	38 – 54%	46%	33 – 54%	5%	2 – 7%	174#
10M: Non-dominant	21%	14 – 30%	21%	15 – 25%	2%	1 – 4%	80
18M: Dominant	35%	25 – 51%	33%	23 – 53%	2%	0.5 – 3%	98
18M: Non-dominant	15%	12 – 21%	15%	11 – 22%	1%	0.8 – 1%	95

Note: # p <.05, adjusted p > .05. The p-values were adjusted using method of Benjamini & Hochberg (1995).

[1] EOS: the sample of every other 30-second segment containing AWCs. Top150: top 150 segments with the highest AWCs. Difference: the difference between proportions of the same type of input in the two samples.

[2] Statistical analysis was not performed for one-on-one input in 18-month dataset to avoid redundancy as the proportion of overheard and 1:1 input added up to 100% in the 18-month dataset. It was not the case for the 10-month dataset because there was a third type of social input, that is group input which we observed none in the 18-month dataset.

10M: 10-month sample; 18M: 18-month sample.

Table 3 *Comparison among Spearman's correlations between infant volubility and bilingual input estimated by different units and sampling methods*[1].

Input Measures[1] / Correlations	Every-other-segment sampling			Top sampling				
				Top150			Top40/20	
	Segment	AWC	Duration	Segment	AWC	Duration	AWC	Duration
10M: Global	.68***	.60**	.62**	-	.49*	.48*	-	-
10M: Overhearing Contexts	.49*	.49*	.52*	–.02	.28	.25	.39	.39
10M: One-on-one Contexts	.50*	.50*	.48*	.05	.07	.11	.32	.32
10M: Dominant Language	.23	.37	.34	.06	.30	.29	.32	.34
10M: Non-dominant Language	.39	.45*	.43	.03	.30	.31	.40	.39
18M: Global	.58*	.49	.49	-	.35	.33	-	-
18M: Overhearing Contexts	.17	.15	.17	–.28	.05	.02	.05	.04
18M: One-on-one Contexts	.44	.57*	. 56*	.28	.45	.46	.58*	.57*
18M: Dominant Language	.50	.61**	.60**	.33	.51*	.53*	.37	.47
18M: Non-dominant Language	.04	.10	.09	–.26	<.01	.01	.09	.07

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

[1] Segment: Segment Count, the number of 30-second segments. AWC: LENA-derived adult word counts. Duration: the sum of LENA-derived Female and Male Speech Duration. Every-other-segment sampling: a sample of every other 30-second segment containing AWCs. Top150: top 150 segments with the highest AWCs. Top 40: top 40 segments with the highest AWCs in one-on-one or overhearing social contexts. Top 20: top 20 segments with the highest AWCs in the dominant or non-dominant language.

10M: 10-month sample; 18M: 18-month sample.

Figures

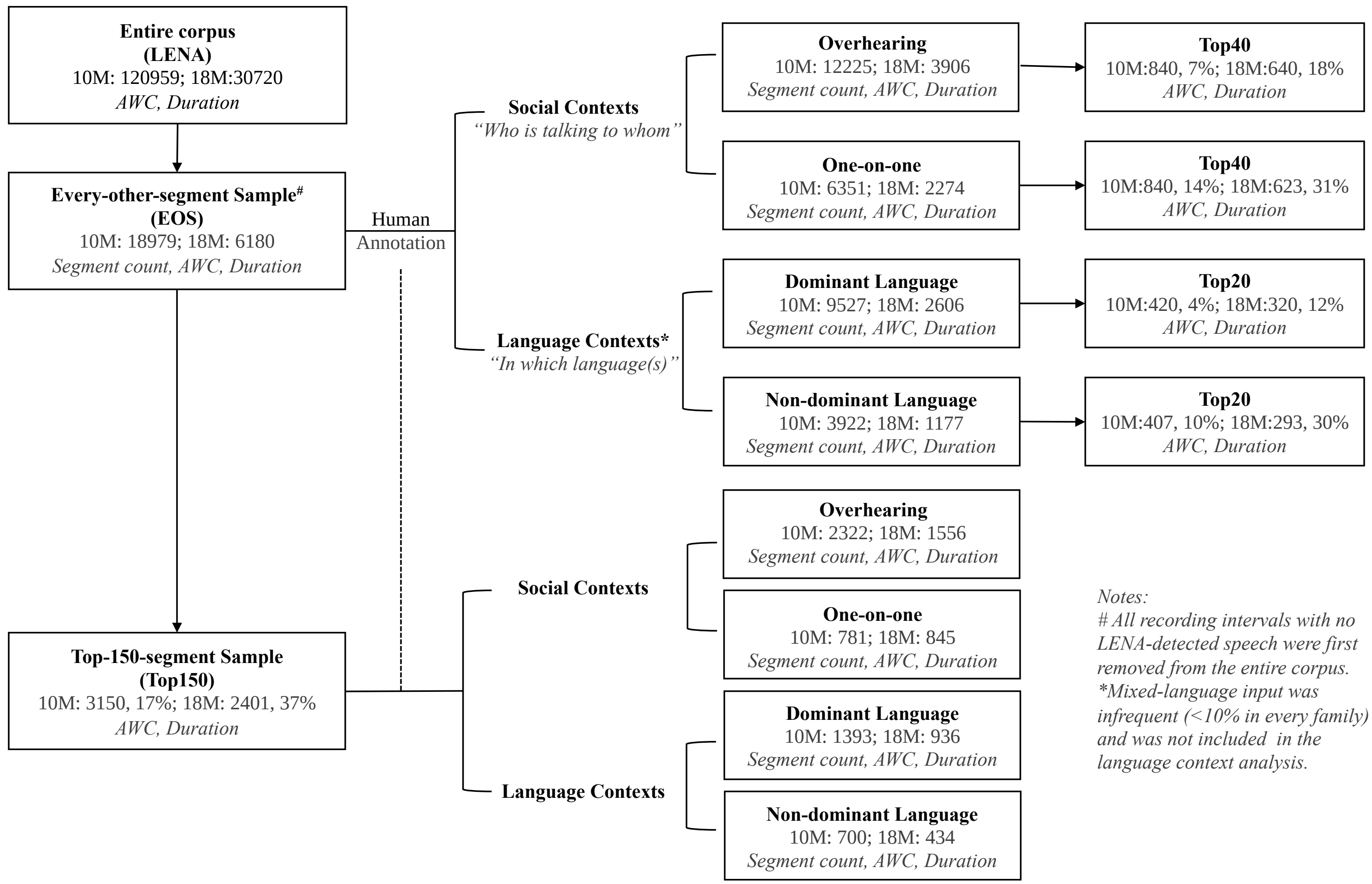
Entire corpus
(LENA)
10M: 120959; 18M:30720
AWC, Duration
Every-other-segment Sample#
(EOS)
10M: 18979; 18M: 6180
Segment count, AWC, Duration
Human
Annotation
Top-150-segment Sample
(Top150)
10M: 3150, 17%; 18M: 2401, 37%
AWC, Duration
Social Contexts
"Who is talking to whom"
Language Contexts*
"In which language(s)"
Overhearing
10M: 12225; 18M: 3906
Segment count, AWC, Duration
One-on-one
10M: 6351; 18M: 2274
Segment count, AWC, Duration
Dominant Language
10M: 9527; 18M: 2606
Segment count, AWC, Duration
Non-dominant Language
10M: 3922; 18M: 1177
Segment count, AWC, Duration
Top40
10M:840, 7%; 18M:640, 18%
AWC, Duration
Top40
10M:840, 14%; 18M:623, 31%
AWC, Duration
Top20
10M:420, 4%; 18M:320, 12%
AWC, Duration
Top20
10M:407, 10%; 18M:293, 30%
AWC, Duration
Social Contexts
Language Contexts
Overhearing
10M: 2322; 18M: 1556
Segment count, AWC, Duration
One-on-one
10M: 781; 18M: 845
Segment count, AWC, Duration
Dominant Language
10M: 1393; 18M: 936
Segment count, AWC, Duration
Non-dominant Language
10M: 700; 18M: 434
Segment count, AWC, Duration
Notes:
All recording intervals with no LENA-detected speech were first removed from the entire corpus.
*Mixed-language input was infrequent (<10% in every family) and was not included in the language context analysis.

Figure 1. Flowchart describing how each sample was derived from the corpus, the number of segments included in each sample, the proportion of segments selected from the original sample (median), and units used to index input in each sample (Italic). 10M: 10-month sample; 18M: 18-month sample.

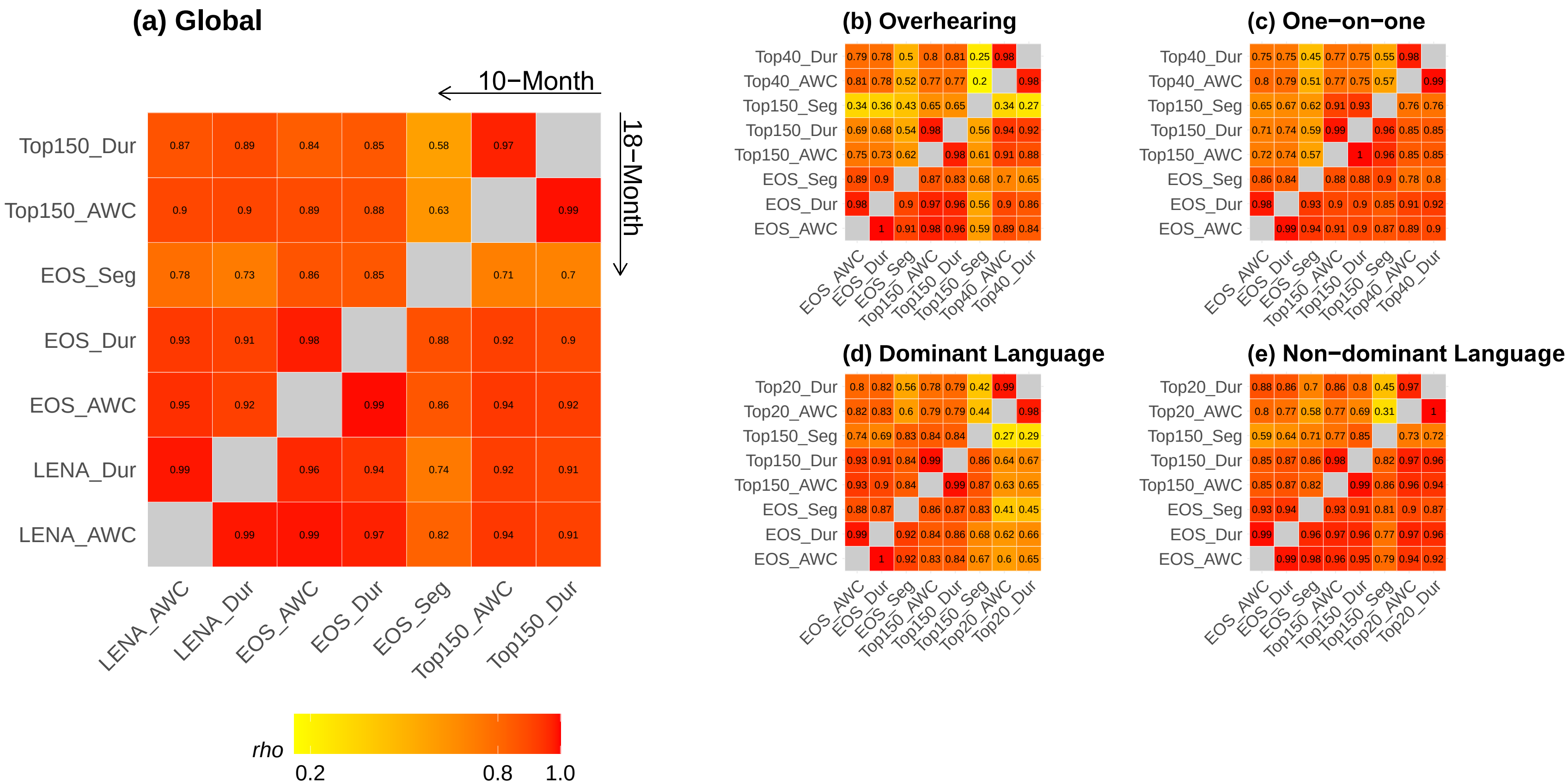

Figure 2. Spearman's correlations between different language input measures in (a) global, (b) overhearing and (c) one-on-one social contexts, as well as (d) dominant and (c) non-dominant language contexts. Upper triangle: 10-month sample; Bottom triangle: 18-month sample. Each cell indicates the correlation between a pair of input measures. The Spearman's *rho* value is reported in each cell. The cell colour indicates the strength of the correlation, from weak (yellow) to strong (red). LENA: the entire corpus. EOS: every-other-segment sample. Top150: top 150 segments with the highest adult word counts (AWCs). Top 40: top 40 segments with the

highest AWCs in one-on-one or overhearing social context. Top 20: top 20 segments with the highest AWCs in the dominant or non-dominant language. AWC: LENA-derived adult word counts. Dur: Duration, the sum of LENA-derived female and male speech duration. Seg: Segment Count, the number of 30-second segments. The columns (C) and rows (R) are referred numerically from left (1) to right, and from bottom (1) to top. For example, [C1, R2] in (a) refers to the cell corresponding to the first column from the *left* (LENA_AWC) and the second row from the *bottom* (LENA_Dur), which shows the correlation between the global input estimated in the entire 10-month corpus by AWC and speech duration. A video-animated guide is available in the Supplementary Material.