Modeling Language Development:
How Machine Learning can Enhance Analysis of the Language Environment

By: James Harvey

1. Introduction

Children acquire language in the presence of complex environmental input that they engage and derive their understanding from. Language environment differences, including those associated with socioeconomic status (SES), result in variation in available information, and are ultimately associated with varied learning outcomes (Anderson et al., 2021). There are positive associations between child language outcomes, input quantity (Jones & Rowland, 2017; Rowe, 2012), and measures of quality including syntactic complexity and vocabulary diversity (Hsu, et al., 2017; Rowe & Snow, 2020). Recording the home in long form naturalistic ways presents an ideal way to sample representatively. In particular, the use of Language Environment Analysis recorders (LENA) presents a way to gather hours of unimpeded exchange in the home. These devices offer the opportunity to sample a far denser amount of language than what would typically be utilized, and as such they present new analysis costs. Before expanding on those analysis costs, and how to overcome them, questions surrounding language sampling must first be understood.

Ongoing questions include how to effectively make use of samples of up to sixteen hours in duration. Studies have relied on shorter samples or taking LENA data at face value, but these each present limitations. Additionally, with those samples of that density, how can analysis draw meaningful conclusions outside of following automatic metrics? Beyond identifying trends in traditional metrics like the number of words, analysis can target the role of activities in home language. More specifically, the use of topic modeling to extract activity contexts from transcriptions of LENA data could present a new lens to assess home language. One concern

with previous methods is that metrics are extracted from the context and language itself, losing many qualitative characteristics relevant for development. This can mistakenly suggest that the quantity of parental input in development is more influential than it actually is, placing the onus of developmental gaps on the amount of speech provided by parents (Grieve, 2021; Hsu, et al. 2017; Huang et al., 2023; Kuchirko, 2019).

In the following sections, I will introduce the study's sampling methodology along with a metric of interest and a novel approach aimed at uncovering that metric. Our data is produced in long form recordings, and we are interested in both the traditional quantitative features of those recordings, but also at ways to identify large-grained context related information from these samples. To that end, we are exploring the role of activity contexts in the home, and using topic modeling as a way to attempt an extraction of these activities from the observable language features.

1.1 LENA Literature and Application

Naturalistic and individually dense recording methods are associated with stronger links between parental input quality/quantity and child language development than shorter, in-lab, or otherwise modified settings (Anderson, et al., 2021; Bornstein, et al., 2002; Cameron-Faulkner et al., 2018). This suggests both that there are relevant discrepancies in the nature of linguistic exchange based on the setting/activity of recording, and that dense in-home recording methods that minimally interrupt the flow of linguistic exchange offer improved representativeness of children's language experience. To accurately make claims about the experience of children learning language, the environment they learn within must be optimally recorded in the form of non-intrusive, individually dense data collection to support the validity of conclusions made.

LENA devices are modified digital audio recorders that can be secured in specially designed shirts to constantly record the surrounding environment of a given child. These devices are designed to alleviate the difficulties associated with sampling densely by making the technology wearable and capable of providing automatic vocal analysis (Cristia, et al., 2021). Although initially designed for parents to monitor their children's language development, these devices have presented a number of applications in research aimed at naturalistic sampling of language acquisition metrics (Cristia, et al., 2021). LENA devices can record up to sixteen hours of audio, and the contained digital language processor (DLP) is able to provide estimates for adult word count, child vocalization count, and conversational turn count along with measures of background noise, meaningful speech, silence, and electronic noise as a percentage of the sample (Cristia, et al., 2020).

Using LENA recorders solves many concerns of previous study designs, but creates new analysis costs and questions of how to use the data. First of all, individually dense data is just that – individually dense. One motivation behind lab manipulations of language sampling is the precision with which a given variable or metric can be observed. LENA devices are not as precise as those lab manipulations and record the varied natural unfolding of the language environment in a way that makes extracting conclusions more complicated. Instead of 50 utterances or a set of questions and responses, the output to be analyzed is sixteen or more hours of environmental noise and language without direction in terms of topics, grammatical features, participants, or quantifiable metrics such as word or utterance count.

Additionally, LENA's automatically calculated metrics are subject to error and do not produce a transcript for the speech contained within samples. For published studies there is a tendency for overestimation of adult word count and child vocalization count along with an

underestimation of child turn counts by the digital language processor when compared to human annotation (Cristia et al., 2020). Fortunately, these imperfect measures can be circumvented, particularly because of recent advancements in automatic speech recognition (ASR) which make the use of these recordings more manageable. One way to avoid taking these values at face value is to use them as an indicator of what segments of a recording to prioritize analysis of.

Although samples of sixteen hours would require up to 176 hours of work by a human transcriber (if sample time is in a 1:10 relationship with manual transcription time), employing new tools for transcribing can alleviate much of the burden on researchers (Gaur, et al., 2016; Novotney & Callison-Burch, 2010). It has been demonstrated that, when ASR is applied with a word error rate below 30%, this step measurably decreases the transcription time (Gaur, et al., 2016). ASR has previously been used unassisted to provide the data for analyzing the home language environment (Greenwood, et al., 2011). For our experiment, the use of a first pass through Whisper transcription software and the pre-delineation of what utterances to retain on a human coded second pass can reduce the time required for transcription from around 1:10 to closer to 1:4. As previously mentioned, in this study, segments of the recording will be transcribed in priority order using LENA's automatic metrics. Cutting transcription costs with Whisper can make an otherwise intractable data set much more manageable and expand the potential usefulness of these recordings.

1.2 Activity Contexts Change Home Language

Even with recordings and their subsequent transcriptions in hand, researchers still have to make several analysis decisions that will shape their conclusions. Understanding the degree to which environmental input predicts language learning as measured by mean length of utterance

(MLU), literacy, vocabulary development, educational success, etc. has long been a focus of research (Hoff, 2003; Lytton, 1971; Murphy, et al., 2022). Research has focused on identifying what variations in environmental input explain gaps in language measures between groups across SES, race, and educational attainment (Betancourt, 2016; Buckingham, et al., 2013; Henry, et al., 2020; Hurt, et al., 2016; Shrifer, et al., 2011). Nevertheless, this work has been limited by the challenge of representing the language environment in its breadth to form meaningful conclusions that relate to the natural unfolding of linguistic exchange. Historically, the solution to this challenge has been either to reduce naturalism in lab settings for the sake of observability or to identify convenience measures and focus on their presentation within given language environments (Hart & Risley, 1995; Lytton, 1971). This work has utilized word counts (Hart & Risley, 1995), conversational turns (Donnelly & Kidd, 2021; Gilkerson et al., 2017), and or type-token-ratio (Hess, et al.,, 1989) as accessible predictive tools for measures of language such as vocabulary development, but these tools may not reflect the natural unfolding of a child's linguistic experience (Baugh 2017; Dudley-Marling & Lucas 2009; Richards, 1987). These are all legitimate features of language, but they may not capture the breadth of what a child is receiving in terms of input. Although the number of words a child hears is a great proxy for how much language they hear, all information about what was said, how it was said, and by whom is lost. These surface level measures are ideal for practicality, but the density of information contained by language is not as practical.

The analysis of activity contexts provides an alternative method for describing the language environment that can avoid some of the pitfalls of other, less rich measures. Activity contexts are the events and interactions in which language is embedded (Holme, et al., 2021). These contexts can be thought of as a latent variable beneath other observable metrics such as the

words used, the participants, and where/when interaction occurs (Roy, et al., 2012). Mealtime, playtime, book reading, or bedtime routines could all be considered activity contexts for a child at home and each of these events are associated with differences in the language used (Holme, et al., 2021). These activities provide an underlying structure that the interaction can be characterized by and that the child can utilize to attach meaning.

In the home, observing activities as they naturally occur supports a representative understanding of what the child's linguistic experience looks like typically. Activities provide the context for language use - context which is often overlooked or manipulated from its natural expression through the course of experimentation and observation. The activity beneath an exchange constrains that exchange across a series of features, and understanding how activities and their effects vary across homes could provide novel insights. The collection and observation of activity contexts can therefore improve the quality of conclusions drawn from naturalistic recordings.

There have already been several examples of studies involving the analysis of linguistic features as mediated by the activity context they are embedded within. These studies have been performed across activities within a given environment (Flynn & Masur, 2007; Ogura, et al., 2006), across languages within activity (Altinkamış, et al., 2014; Doering, et al., 2020; Glas, et al., 2018), across SES to identify patterns of activity presentation (Hoff-Ginsberg, 1991; Rosemberg, et al., 2023), within activity with varied activity complexity (Muhinyi & Hesketh, 2017), and within parent-child interaction as a motivator for linguistic features (Crain-Thoreson, et al., 2001; Soderstrom & Wittebolle, 2013). Additional studies have taken a longitudinal approach and attempted to identify linkages between the activity contexts observed in the home and future measurable language outcomes, either through isolating these activities and their

presentation between families (Demir-Lira, et al., 2019; Gilkerson, et al., 2017) or through the persistent recording of a single language environment supported by location and time information (Roy, et al., 2012). These studies have demonstrated links between activities and the observable linguistic exchange (Holme, et al., 2021; Ogura, et al., 2006), suggesting that activities motivate certain language behaviors even when comparing across languages (Doering, et al. 2020; Glas, et al., 2018) or complexities within the same activity (Muhinyi & Hesketh, 2017). Considering the role of activity contexts in patterning linguistic exchange, the identification of these contexts presents an enticing opportunity for insight into the language environment, particularly with naturalistic recording methods. A key question that arises from prior research is how to best record and then identify these activities in their natural presentation, and how to compare when they occur for different families. Considering that we want to understand activities as a latent variable that constrains features of language including which words are used, adapting natural language processing techniques to this data set could yield new insight.

1.3 Topic Modeling for Dense Data

Recovering meaningful information about communicative activities from natural language use is a challenge. Natural language refers to any human language whether written or spoken as opposed to a language formatted for computers to understand. Natural language is full of superfluous features that may not connect to the meaning when compared to programming languages. To automate analysis of natural language, context is often removed in favor of quantifying surface-level features, but this overlooks the importance of context in setting the

stage for the nature of linguistic exchange. To gather information about context, tools must be employed that are capable of analyzing dense natural language data.

One approach to this end is the use of probabilistic topic modeling, a method for analyzing dense data popular in natural language processing. Probabilistic topic modeling is a statistical machine learning technique that can be used to organize and search across large-scale datasets to produce salient themes from within a given sample (Blei, 2012). Latent dirichlet allocation (LDA) modeling is particularly useful for analyzing unstructured or highly variable data as it functionally "zooms out" of the data to focus on the themes and interaction between themes rather than particular features (Blei, 2012). The process of LDA topic modeling acts without regard to the order or meaning of the input, and instead focuses on metrics about the occurrence of words (Kherwa & Bansal, 2019). In this way, the model processes the input as a bag of words: an ungrouped set of tokens across the sample rather than an organized continuous document. This methodology allows easily permutable scaled data analysis across the transcribed LENA recordings (Kherwa & Bansal, 2019). The creation of these models is motivated by a number of assumptions about how given observable features of language come to be.

The primary assumption is that all texts and samples, referred to as a collection of documents moving forward, are formulated through a process that combines observed and hidden variables to generate their realized form. An observable feature would be any words retained from the sample and hidden variables motivate the form of that observable input. In topic modeling, topics are assumed to consist of a set of similar words within documents. When active, a topic will make the use of certain words more likely relative to the frequency of that word in the input corpus at large. In this way, a topic within topic modeling could function similarly to an activity in the home. Each of these features motivates variation in the observable

home input, but is invisible within the input itself. Variation between activities or topics would generate variation in the frequency of tokens across a given sample. Understanding how a document takes its form motivates the way that topics are extracted.

For example, before writing a newspaper article about a restaurant and their cuisine, a journalist may consider the story they want to tell and the beats they need to include along the way. To transform this story from ideas into an observable article, the author must use sufficient detail to convey their points, must follow the grammatical rules of their writing language, and do so within a certain amount of words or space on a page. Topic modeling uses statistical analysis to uncover the ideas that generated that writing. For a person to manually identify topics in an article, they would have to read through each section and try to identify what the key features of the article are. In doing so, they would generalize across the article, ignoring much of the observed form in an attempt to identify what topics are behind the article. Topic models use statistical analysis motivated by the assumption about how words are embedded within topics across documents, to automate this same process of identifying hidden variables.

All words within the input corpus are made into a bag of words and then a document term matrix where each unique type is stored with an associated frequency count. From there, the modeler asks for hyperparameter inputs on the number of topics to extract and the number of iterations to run. Then each word within the document term matrix will randomly be assigned to a topic, and each document will be randomly assigned a proportion of each topic. The model uses these random distributions and hyperparameters to recreate an output corpus of identical volume. When the modeler does a good job at identifying the relationships between words and topics and topics and documents, the resulting output will look similar to the input. Whichever iteration most directly reflects the input has the most accurate identification of topics. That

winning model will be presented in the form of unlabeled topic lists that require human interpretation.

1.4 The Current Study

At its core, this study is concerned with sampling and analysis metrics. Sampling is the way of getting ground truth into the form of analyzable meaningful data, but every sampling choice manipulates the conclusions drawn. Similarly, analysis is a necessary abstraction, but one that comes with associated costs and benefits. This study is evaluating the efficacy of a pipeline of producing, transcribing, and analyzing long form naturalistic samples using LENA recorders. To validate that this method of sampling produces useful data salient to the nature of home exchange, sampling will be validated by scanning for expected effects of socioeconomic status and parent input on home language and child performance. The presence of expected trends would suggest that this unconventional method creates representative data, and it would provide further support for this pipeline's application beyond associated time benefits.

To validate our sampling approach, the data will be evaluated for SES and developmental effects. Mixed effects models will consider the effect of participant income and child vocabulary size on the observed language features while controlling for participant level variation. If this sampling method is representative and recreates expected patterns, then families of higher SES and children with higher standardized PPVT scores are expected to have a greater number of words (NOW), number of different words (NDW), and mean length of utterance in words MLU-W. The effects of this SES variation on language would also be expected to be less strong than the effect of parent language on child vocabulary scores on the PPVT. If this method is not

valid, then the patterns in variation may be in the opposite direction or otherwise missing altogether.

In the second part, we will implement topic modeling as a novel analysis of home language. This data is sparse relative to typical topic modeling applications, but it is dense compared to typical language samples that a speech pathologist would analyze. This density of language presents costs at the analysis stage due to its sheer volume and the difficulty of selecting and extracting meaningful conclusions. To overcome these density costs, topic modeling will be applied in the hopes that it can uncover the latent variable of activity contexts within the data. This will be evaluated first by an attempt at human annotation of activities from the topic lists, and then by observation of an inter-topic distance chart that allows the model to be evaluated in a 3D representation of linguistic space to identify if the lists seem distinct.

Completing this study will give insight into the nature of home language exchange, sampling, and analysis. If expected patterns are uncovered in the form of t-tests and mixed effects models, then the sampling will be validated as capturing the home effectively. If this pipeline produces valid data, then it should be shared as a way to scale up sampling while still improving efficiency. Once transcripts are in hand with sampling validated, topic modeling presents a new mode of exploring the data. If effective, topic modeling could provide an interesting supplemental analysis to traditional language sampling metrics.

2. Methods

2.1 Subjects

This project involved secondary analysis of data collected from 32 families in the Washington, DC metro from varying SES backgrounds. Based on a median split, SES

background was split into two groups (low and high) in reference to the comparison of the family's yearly income to 70-100k where a family with an income over this range was labeled "High SES" and a family below this range was labeled "Low SES". Within these delineations, there were 20 participants from low SES backgrounds and 12 from high SES backgrounds. Children varied from 4;0-7;5 on the days of recordings with a median age in the range of 5;0-5;5. Nine of the participants were recruited from Howard County Community Action Center Head Starts, and the remaining twenty-three were from other parts of the DC/Maryland area.

2.2 Materials and Procedures

Participating families were mailed the LENA device and accompanying shirt and asked to record an eight-hour sample on two different days for sixteen total recorded hours. Upon return, recordings were uploaded to LENA Online. By default, the recordings are broken into five-minute segments. LENA's proprietary algorithms automatically calculated the metrics mentioned above (AWC, CVC, CTC) per five minute segments of the sixteen hours. With these metrics recorded and accessible in a CSV, a Python script was implemented to sort from most to least "meaningful speech". Meaningful speech was identified by the total speech coming from nearby human sources within each segment. This process identified and ranked segments by the amount of meaningful human speech to reduce the analysis of samples laden with silence, background, or electronic noise.

Although the estimates that LENA provides are subject to error, because we are transcribing the segments, we are not taking their figures at face value. Instead, an automated process hurries the analysis by giving a starting point for which segments of the total sixteen hours to analyze. LENA's metrics and Whisper's transcripts are simply the initial place from

which human analysis will eventually interfere to ensure precision and accuracy. Using the volume of meaningful speech to prioritize segments should allow samples to include the most dense parts of the day in terms of language exchange.

These sorted five minute audio clips were automatically transcribed with Whisper, OpenAI's AI-assisted automatic speech recognition (ASR) system ("Introducing Whisper.", 2022). This process outputted transcriptions of all words found within the audio file segments with higher accuracy for adult speech ("Introducing Whisper.", 2022). To improve this first pass transcription, human research assistants examined each text file while listening to the segment's associated audio file to ensure the transcripts were accurate. In addition to fixing up mistranscriptions, transcribers also separated the available adult utterances while removing any transcribed child speech, background, and electronic noise.

Electronic noise, for the purpose of this experiment, included any captured TV, phone, or otherwise virtual speech that was not directly present or interacting with the child. This allowed for some cases, such as a participant talking to their parent on a video call, to still be admitted as a part of the language environment without giving the same distinction to remote speech not directed at the child. Once recordings were transcribed through the two-pass approach, the "cleaned-up" transcripts were concatenated into a single file for each participant. Analysis was run at the segment level to ensure that the sampling method reproduces expected variations across SES and parent input. The concatenated files for each participant are the input for topic modeling.

.

3. Results

3.1 Validating Language Samples Using Traditional Metrics

The first goal of this study is an evaluation of a new sampling methodology. This study is not the first to use LENA recorders to extract information about the home language environment, but the pipeline that the data goes through is relatively novel. Because this is an unconventional approach, the data cannot be taken at face value and must be validated by comparison to evidence-based patterns in variation. Before analysis is conducted on the language features captured by this methodology, early results must be compared to established findings. The two findings that serve as validation for sampling have to do with the effect of SES on language and the effect of parent language on child vocabulary.

The relationship between socioeconomic status and language has been explored from many angles, but the most relevant for this study is the effects of socioeconomic status on parent language and on natural recordings of home language. If expected trends are recreated, parents from higher socioeconomic status backgrounds are expected to have a greater NOW, NDW, and MLU-W, but these effects are not very strong in naturally recorded home language (Huttenlocher, et al., 2010; Piot, Havron, & Cristia, 2022). The other primary finding that this study is seeking to replicate is the effect of adult language on child vocabulary and standardized testing performance. Particularly, vocabulary size is related to parental input and this study will directly assess the effects of parental input on child vocabulary performance (Arriaga, et al., 1998).

To assess whether expected trends are retained within this sampling method, first t-tests were run to assess the relatedness of SES and each of the primary metrics of interest (MLU-W, NOW, NDW, and number of utterances). The results of those analyses are seen in Figure #1. The

expectation is that there will be a greater NOW, NDW, MLU-W, and number of utterances in homes of higher SES compared to lower SES. Therefore, if LENA is sampling in an effective manner, then the data should reflect that established trend. When observing the data like this, cut along a median line rather than continuously, it becomes clear that all variables of interest, except for number of utterances, demonstrate this trend. There are statistically significant variations in each of these language metrics when comparing across SES. Mixed effects models looked at this same relationship of SES and language metrics.

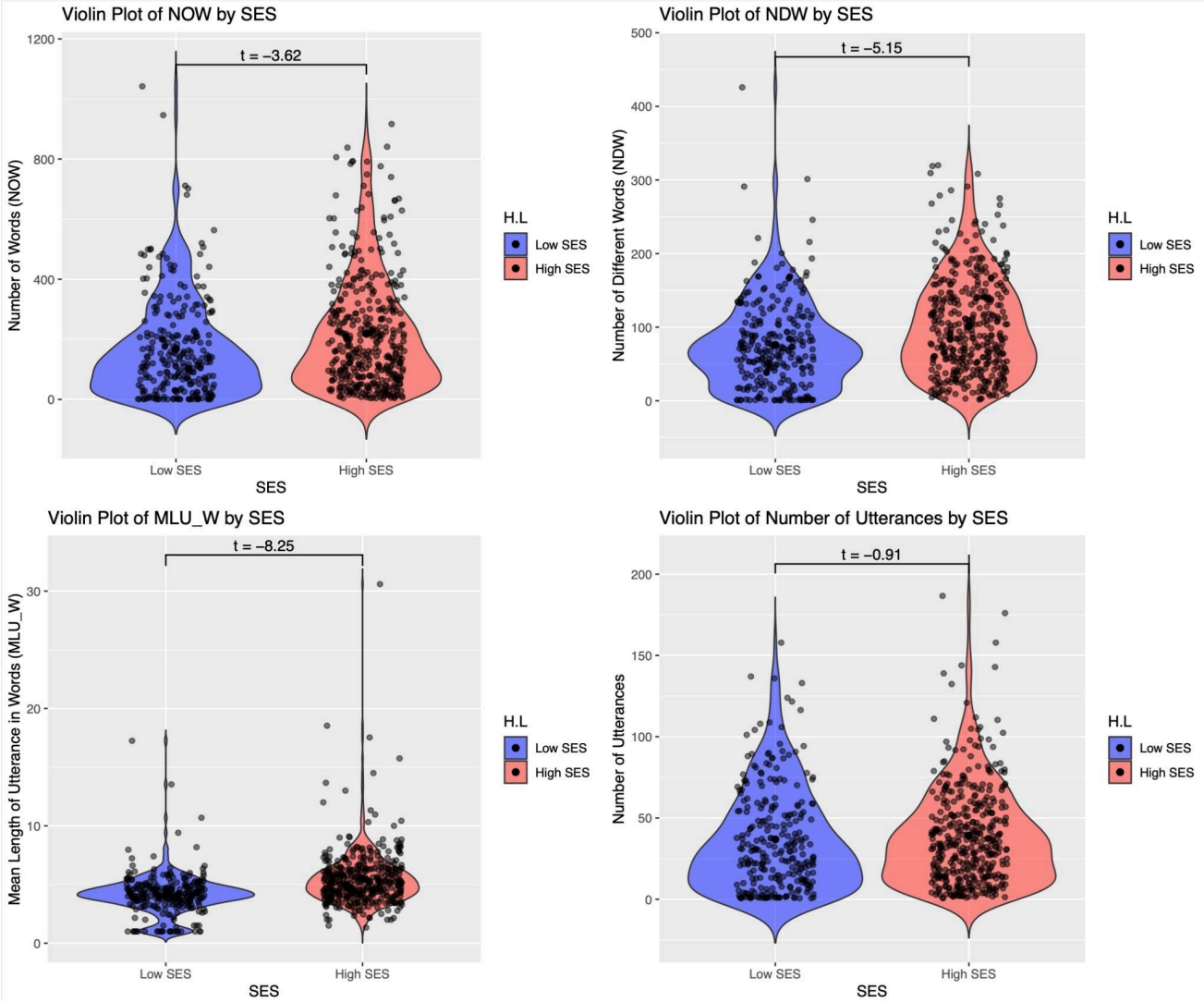


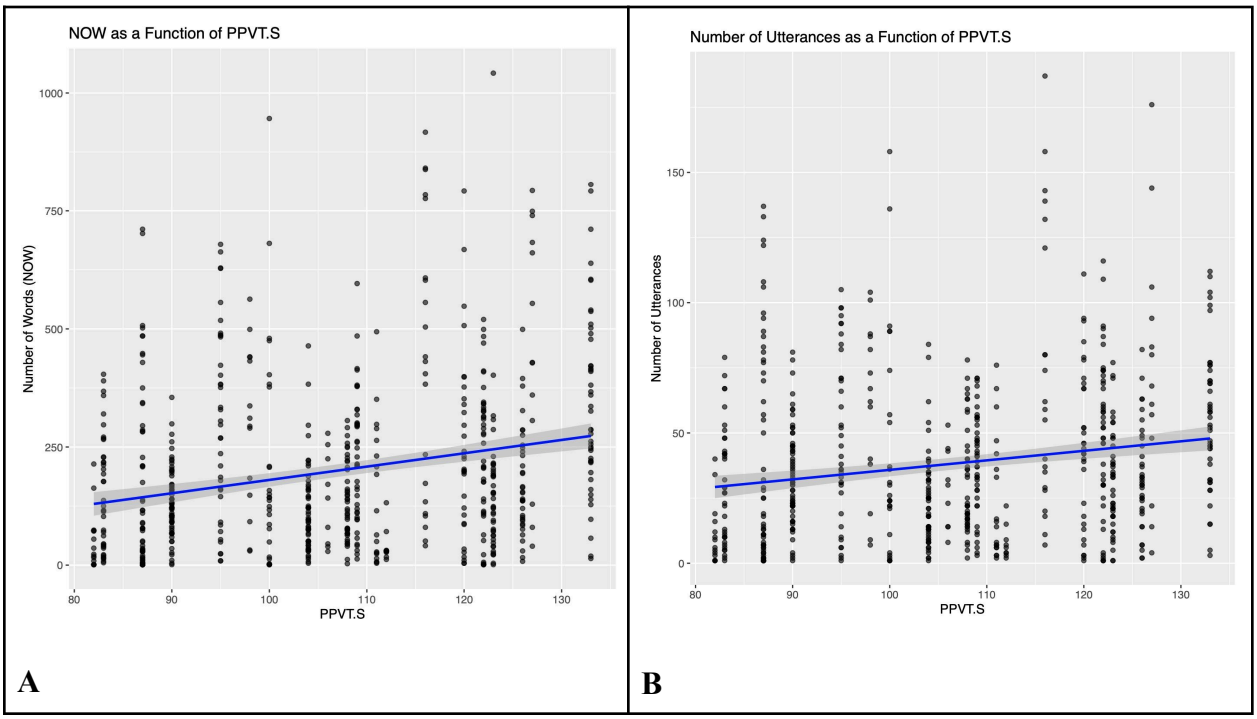
Figure 1: SES T-Test results

Each panel tracks the relationship of a language metric and home SES separated into high or low SES. A: Number of words (NOW), B: Number of different words (NDW), C: Mean length of utterance (in words) (MLU-W), D: Number of utterances.

Mixed effects models allow statistical disentanglement of multiple mediating variables. In this case, variation in the recorded home language is expected to be explained by participant level variation and as the result of fixed socioeconomic status effects. To assess what the effect of SES is on home language, mixed effects models must include the random effect of participant level variation and analyze the fixed effect of SES on the dependent language metric variables. Within these mixed effect models, the only statistically significant relationship is between SES and MLU-W ($t = -4.997$, $p = 0.000114$). All other comparisons were not significant (p's > .05). Although t-values demonstrated the significance of differences in these features, mixed effects models suggest that MLU-W is the only feature that is tied in particular to SES whereas the others may be more impacted by participant level variation. The next validation has less to do with expectations about features of the home, and more to do with the effects of home variation.

For the second set of validation efforts, the linkage between parent input and child understanding is investigated through the Peabody Picture Vocabulary Test (PPVT). Mixed effects models were once again built with participants as a random effect. In these models, child PPVT standard score is plotted by the home language metrics. This analysis is directly investigating how predictive PPVT score is for the features of home language. The results of that analysis are in Figure #2. Within this chart, each point represents a five minute segment from a participant of a given PPVT score as plotted on the x-axis. Each of these variables demonstrate a positive correlation, but most of the metrics have a weak to very weak correlation between the language metric and PPVT performance. The positive direction of the relationship was as

hypothesized, but to understand the relationship between these variables better, mixed effects model results must be considered. There are statistically significant relationships between PPVT score and MLU-W ($t = 2.733$, $p = 0.0135$), NDW ($t = 2.99$, $p = 0.0079$), and NOW ($t = 2.38$, $p = 0.0285$) when controlling for participant level variation. This suggests that there is a stronger relationship between the features of home language and a child's understanding (as measured by PPVT) than the relationship between parent SES and home language.



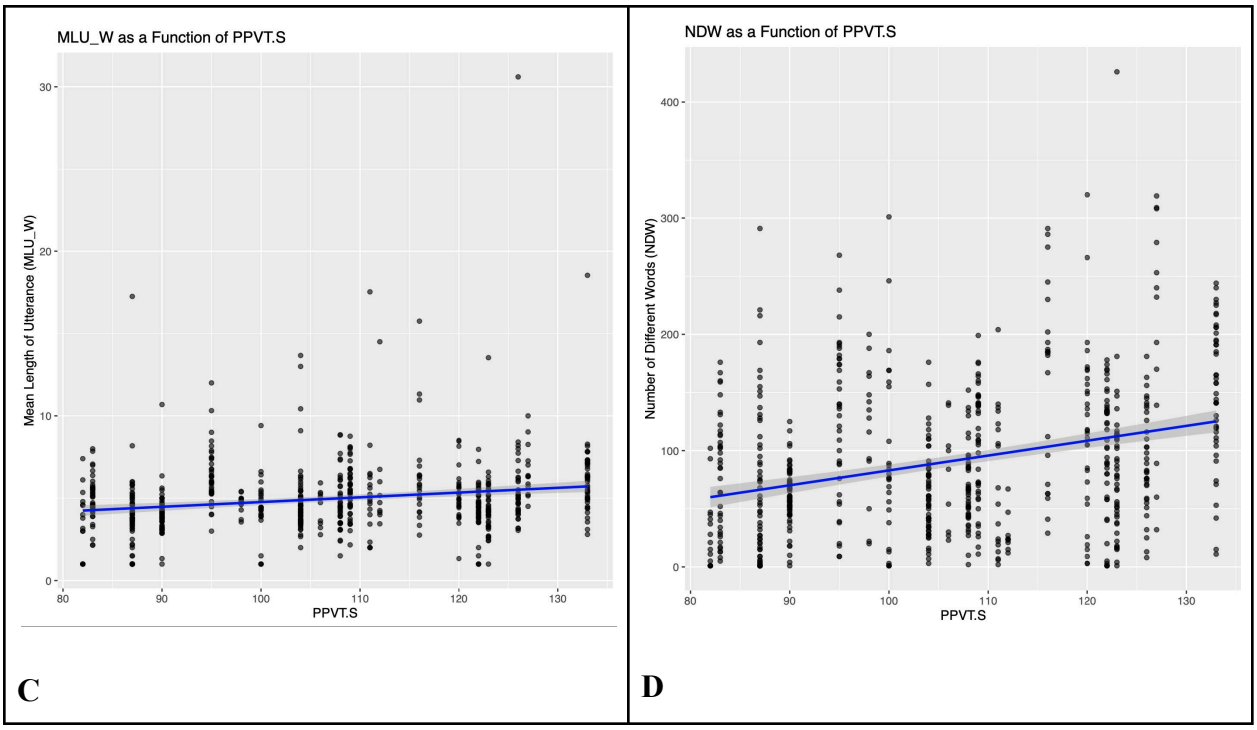


Figure 2: PPVT Mixed Effects Model Results

Each panel tracks the relationship of a language metric and child PPVT standard score. A: Number of words, B: Number of utterance, C: MLU-W, D: NDW. There are statistically significant relationships between PPVT score and MLU-W ($t = 2.733$, $p = 0.0135$), NDW ($t = 2.99$, $p = 0.0079$), and NOW ($t = 2.38$, $p = 0.0285$)

. Thus far, the sampling methodology has recreated the expected trends in home language across SES and in children's performance across adult input conditions. These mixed effects models along with the T-tests all suggest an underlying pattern of both SES predicting home language and parent language predicting child vocabulary to a greater degree. There is a lesser predictive strength of SES for parent language than there is for parent language and child PPVT score. This was the expectation considering the naturalistic data collection method tends to show weaker SES effects on home language (Piot, Havron, & Cristia, 2022). Because these trends are extractable from this sampling methodology, the pipeline from LENA distribution to transcript analysis seems to be a valid way of sampling the home.

3.2 Topic modeling

Before topic lists are extracted from participants' concatenated transcripts, decisions must be made in terms of text preprocessing and the number of topics to output. In preprocessing, input text is aligned closer to the meaning of the words contained within the sample through processes like stop word removal. The removal of stop words, words which carry grammatical function without an effect on meaning, is a crucial step to producing meaningful conclusions. This reflects the analysis goals of the topic modeler in identifying the hidden meaning-related variables beneath the observed features. For the purposes of this project, 127 grammatical stop words were removed using the words identified within the Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009). These words would not saliently contribute to a meaningful topic, or in this case activity context. Words like “the” do not have the contextual weight associated with a distinct topic or activity. Including those words would introduce more noise into the system and lead to less interpretable topic lists.

Beyond removing stop words, the topic modeler also requires input for the number of topics to identify. For our project, the modeler will produce five topics per transcript. Within our data, this decision produced more coherent topics compared to larger or smaller quantities as measured by the LL/token metric and comprised a large enough set for interpretation. LL/token is a measure of a model's log-likelihood over the total number of tokens where a higher value is associated with better model performance. Topic lists greater than five tended to produce a lower LL/token when applied to our data, so five topics was the ideal hyperparameter by that metric. We tested the model’s success producing lists of three, five, ten, and fifteen topics for two participants, and this led to our decision. Additionally, because topic modeling requires human

interpretation, reducing the number of topics to increase LL/token would only make the available information more limited in utility. Maintaining five topics allowed for a balance between human interpretability and statistical support in the form of LL/token. All data preprocessing steps were evaluated by varying the hyperparameters of the number of topics, stemming, and stop word removal and observing the outcome. The final version of each transcript was processed using the Gensim Latent Dirichlet allocation (LDA) topic modeling algorithm.

Topic model evaluation is a challenging task that is still being perfected, but it requires a set of statistical and human interpretations. The first pass evaluation is the human interpretation of topic lists. For our study, when reading through the identified words, is there a salient theme that suggests the model found an activity context? The rest of the evaluation of the output is statistical in nature. The model outputs values for the LL/ratio and the weight of a topic within a document. Secondary analysis, in the form of intertopic distance charts, depict the distinctness and scale of topics within the corpus in the form of clusters. These visualizations give insight into how separate and weighted each topic is, and overlapping topics are considered less salient or meaningfully different from one another. Success for the model would be the production of a set of topics that a person can read and then identify as belonging to an activity. This success would have to be supported through statistical means as well. Failure for the model could be topic lists that cannot be identified as belonging to any activity or an inability to find hyperparameters that work across participants. These models are necessarily swayed by the hyperparameters and preprocessing, so initial failures are expected in the process of refining the pipeline. For topic modeling to be considered a success however, the user-inputted hyperparameters and text preprocessing must be able to remain static while generating meaningful topic lists across participants.

The goal of finding a set of preprocessing decisions and hyperparameter adjustments that would uncover activity contexts from dense unstructured transcripts with topic modeling was unsuccessful. The modeler's performance is considered a failure on account of an inability to generate coherent activity related topic lists. This can be considered a first pass failure because without reference to LL/ratio or any other evaluative metric, the topics are not coherently interpretable to begin with. This means that even if the model rates its performance highly or if there were distinct clusters when observing the inter-topic distance, the process still failed to identify activity contexts. This failure occurred both across and within participants as topic lists contained words with low contextual weight that could not be reasonably interpreted as constrained to an identifiable activity context. Analysis of LL/token and intertopic distance are not strictly necessary to evaluate the effectiveness of this pipeline, but intertopic distance charts were created nonetheless.

Figure #3 is an example of the intertopic distance chart of one participant's data. This intertopic distance chart has more distinct topics relative to other participants, but observing the size of these topics reveals a different problem. While these topics are not on top of one another which would suggest that they are not distinct in tokens, they are generally small with one exception. This indicates that within the input, those topics make up a small percentage of the overall data set relative to a larger cluster. An intertopic distance chart alone is not sufficient to evaluate the identification of activity contexts, but when the chart displays a large amount of overlap or very small clusters, then it would appear that the model is not working effectively.

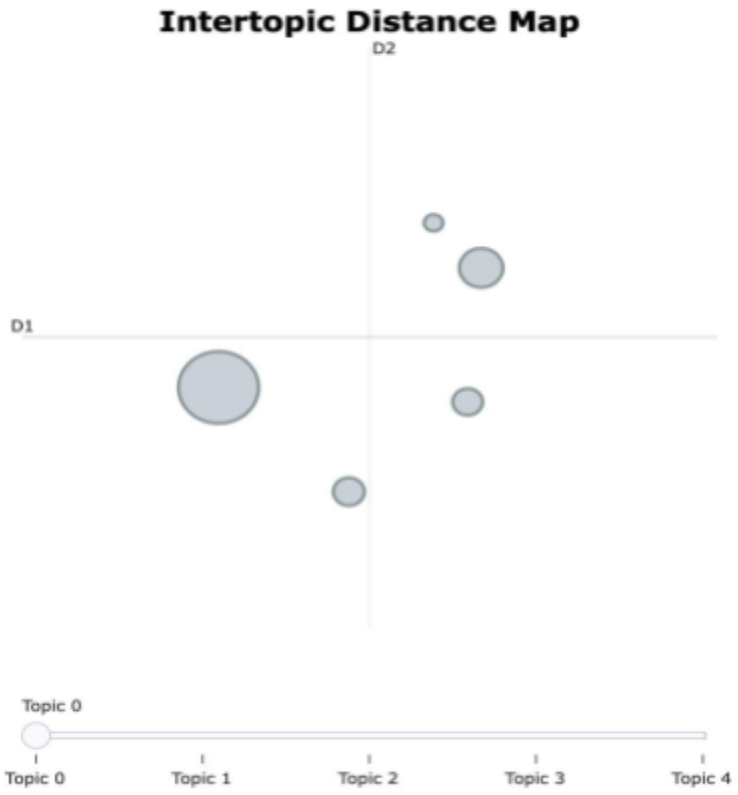


Figure 3: 37_SF Intertopic Distance Chart

Pictured above is one intertopic distance chart from participant 37_SF. Within these charts, the size of the circle represents the weight of that topic within the corpus. The location represents the parts of the corpus that the topic includes, and distinctness suggests that topics consist of unique parts of the corpus.

4. Discussion

This study assessed a new method of sampling language that attempted to solve some of the analysis concerns associated with LENA recorders while still taking advantage of their naturalistic dense recording style. After sorting segments according to meaningfulness of speech and transcribing with a technology-assisted two-pass approach, the sampling was validated by t-tests and mixed effects models that targeted both the effects of family SES on home language and the effects of home language metrics on child performance on PPVT tests. The results of these statistical analyses recreated effects found by other methodologies which serves to validate

the data's pipeline in producing a representative picture of the home language. In combination with a novel transcription approach, the study also assessed the effectiveness of topic modeling at extracting themes in underlying activity contexts from home language. This goal was unsuccessful, likely as the result of an incompatibility between the data and topic modeling.

The following sections will cover the results of the mixed effects models and potential explanations for why topic modeling failed. The mixed effects models justified the sampling method and there are many potential reasons why topic modeling could not extract activity contexts. Finally, the limitations of the study and potential future directions will be discussed.

4.2 Evaluating Mixed Effects Models

Mixed effects models were constructed with and without segments that contained no utterances. Where the previous analysis provided insight into the home language only including segments with speech, the results change greatly if empty segments are included. Observing the model results when all segments are included, there are significant effects of the number of words and number of different words on a child's PPVT performance. The number of utterances and MLU-W are not significant predictors of PPVT performance when segments without utterances are included in analysis. When only segments with speech are analyzed, MLU-W, NDW, and NOW are all significant predictors of PPVT standard scores. These results suggest that all three of those metrics mediate the relationship between a child's innate ability and their observed performance on language tasks.

Beyond finding a meaningful connection that validates the sampling method, this points to how crucial each analysis decision is in shaping the results of a study. One of this study's underlying analysis decisions was the choice to transcribe using a two-pass approach. This left

some of our sampled segments without any utterances. For our purposes we decided to report the results only including segments that had retained adult speech, and the rationale for that decision is found below in the limitations section.

Automated transcription allowed for scaling up of analysis without time or accuracy costs. Researchers are not active in the process of recording the sample or completing the first pass of the transcription. This means that a researcher could feasibly have a first pass of a sixteen hour sample with audio files and transcripts in as long as it took their computer to process the files. Of course, there are also costs associated with sending out and waiting on the return and upload of LENA files, but these recordings can be created parallel to one another and parallel to the analysis of other samples making this less of a factor.

Transcription is a highly costly part of the process of language sampling typically. Developments in artificial intelligence have made it possible to more accurately produce automatic transcripts. Even where these systems make transcription errors they should be caught by the research assistant responsible for trimming the transcription. For unambiguous words, the transcriber should be able to quickly affirm Whisper's transcription, but the ambiguous cases present a more interesting scenario. In the event that a speech sound is difficult to discern, having a first attempt at a transcription could bias transcribers to hearing something that may not really be there. Early research into the quality of data produced by this transcription pipeline is highly positive. When comparing segments where a research assistant transcribed without a first pass to those same segments transcribed with the two-pass approach, the transcriptions were within the goals for inter-transcriber reliability. Beyond accuracy, the time gains of this system are clear.

Another primary benefit of this approach is that Whisper can be run locally which means the potentially sensitive data being transcribed will not have to be uploaded to an unsafe server

or cloud. This two-pass transcription produces equally accurate results faster and just as safe as traditional methods. Primary drawbacks of this system are the technological savvy necessary to run the programs, the still limited ability for these systems to transcribe child speech effectively, and the related troubles extending automatic speech recognition to multilingual and clinical populations. The technological barrier could prevent people who would benefit from this approach from utilizing it. As ASR improves, the accuracy difference between adult and child speech will hopefully reduce, but the specialized language of clinical populations may lag behind further. Additionally, if a family or individual switches between language or dialect, these systems are less attuned than a knowledgeable transcriber. Two-pass transcription is not a universal solution, and it is optimized for cases of large scale adult language transcription, but in terms of efficiency, accuracy, and safety, two-pass transcription matches or exceeds what can be achieved by raw transcription.

4.2 Why Topic Modeling Failed

This study evaluated the extent to which topic modeling can extract salient activity contexts from long-form natural language sampling. The unlabeled topics produced by the topic modeler could contain tokens embedded within identifiable activities. If the process of transcription, preprocessing, and topic modeling can provide insight into the activity contexts presented within the 16-hour recordings, this automated process can be advantageous for gaining insight into the naturalistic presentation of activities across homes and across SES.

Our hypothesis was that the topic modeler will be able to identify these activities when analyzing the transcripts of adult utterances. This hypothesis is supported by research that suggests activities mediate the features of linguistic exchange including both the choice of words

and utterance complexity even across different language environments (Holme, et al., 2021; Ogura, et al., 2006; Rosemberg, et al., 2023). This hypothesis is further supported by the value of topic modeling for analyzing dense unstructured data, and the assumption that the output of these models will be markedly different for home language than for written word. LENA data is not structured like a series of structured written documents, it is not centered around topics, nor does speech follow the same degree of grammatical rigor as writing. These factors produce documents much different from the typical application of topic modelers, and this variation may be advantageous for the automated identification of activity contexts.

There is already existing work that utilizes machine learning techniques such as topic modeling to identify activity contexts in natural language (Roy, et al., 2012; Wang, et al., 2014) Although these studies both dealt with denser data, and Roy's work included measures of where and when language was exchanged, these studies function as benchmarks for the successful application of topic modeling with the objective of activity context identification. Our study presents a trimmed down data set with the hopes that the function of the topic modeler over a smaller body of text will still be able to similarly extract meaningful information from natural language exchange.

Topic modeling is optimized to run over a large input corpus to find patterns in word distributions that are artifacts of some hidden "topic". In our case, the goal of topic modeling and the nature of the input are different. We neither want to identify the topics of conversation in the home (e.g.: sports, work, school), nor do we have data that was written or produced with an explicit purpose. This project acts under the assumption that activity contexts are a latent variable that can explain some amount of observable speech output. This is similar to the assumptions that make topic modeling effective over written material produced with themes

"behind" the analyzed documents. The difference is that these activities are not conceptually related in the same way to the output as those themes are. The themes that a journalist includes in an article are *built* by the words that a topic modeler may identify, but an activity is underlying and maybe only makes certain words more likely relative to their occurrence across contexts.

The reasons for this failure may lie in the nature of the data and or in the way that activity contexts constrain language differently from a topic. While samples of sixteen hours are large relative to typical clinical or research language samples, they pale in comparison to the scale of typical applications of topic modeling. Although the nature of the segment selection and transcription process results in unstructured individually dense data, it may be the case that there are simply too few documents for topics to become coherent beyond commonly used words. Alternatively, the data may never be appropriate regardless of density simply due to a data and analysis misalignment.

Topic modeling was developed to extract themes from large sets of written documents. The relationship between topic and observable output in that case is drastically different from home language. In an article, the author has some concept or topic that they materialize in accordance with stylistic and grammatical limitations. In the home, and in conversation in general, this process is significantly messier. First, topics of discussion are not predetermined or clearly distinct and the flow of a discussion is impacted by all participants and their perceptions. Second, language is a means to an end in a different way in conversation than in writing. Where an article must be able to stand on its own in communicating some set of information, communication in person can rely on visual cues, shared understanding between participants, and other pragmatic forces that would otherwise escape transcription. As such, the features that shape observable output are more complicated and varied in home natural language than in written

documents. The tool of topic modeling was designed for a certain goal, and the contrasts between its intended application and the current study's data could be too great to reconcile. Failure could also be the result of the way activity contexts interact with language.

Although there are documented effects of the activity context on the nature of linguistic exchange in the home, those effects may not be clear in this sampling methodology. Parents are asked to record two "typical" days that the child is not in school for. These recordings therefore contain all of the activities of that child's two days, but topic modeling may be blind to the effect of these activities.

First, only a portion of the segments are analyzed, namely the segments expected to contain the most speech. Therefore, if an activity were to increase the sheer volume of language, it would also be disproportionately represented in the concatenated transcripts. Secondly, the topic modeler is aggregating across the provided text input, input that lacks child speech which may have further contributed to the salience of activity contexts. Lastly, similar to how the data may be too sparse for topic modeling, it may be too sparse for the effect of activities to be visible from only the adult speech of two non-school days. There may not be a large enough range of densely represented activities within that time, and the effects of activity variation may therefore be lost. Perhaps encouraging parents to complete some of a set of preexisting activities during their recording could introduce greater variety in a controlled manner. Fortunately, this study was primarily focused on evaluating the sampling methodology's representativeness.

4.3 Limitations & Future Directions

This study's primary aim was an evaluation of a sampling and analysis methodology. Part of that validation process involved comparisons across socioeconomic groups and an evaluation

of the effects of that participant characteristic on the observed features of home language. T-tests and mixed effects models generated evidence that the expected variations are recorded within this sampling method. Unfortunately, although our sample includes families across SES, our LENA data from that sample is less evenly distributed. Within the recruited families for the project that this study is nested in, parents from lower SES households were more hesitant to allow the recording of sixteen hours of their children's lives.

There are many explanations for why someone may not be comfortable with this recording method, but parents cited specifically that they may not have a regular enough routine across non-school days, or they are in a single parent or separated household where they aren't always around their child. Research has also consistently been extractive and pushed deficit models that would not encourage participation by members of these groups. The primary effects of this hesitance are twofold. First, for the purpose of the mixed effects models, SES had to be considered as a categorical variable along that $105,000 annual income line. Second, there is simply an overrepresentation of the language environments of higher SES families. This may limit the study's ability to make sweeping conclusions about home language variation in relation to socioeconomic status. Beyond validation of the sampling, segment selection decisions may have limited the salience of activity context effects.

The feature of language that we use to select segments varies as a function of activity and therefore biases our sample towards the activities that contain the most speech. More specifically, we sorted the segments based on which were labeled as having the most meaningful speech. If during playtime, parents spend more time talking to their children than they do during other parts of the day, then segments recorded during playtime would be expected to make up a greater proportion of the analysis than the sample itself. This would therefore reduce the

representativeness of activities as they actually occur. To circumvent this, random sampling of segments would be more effective if activity context identification is the primary goal. Segments were also somewhat hindered by LENA's tendency to overestimate the amount of speech in a given segment.

Our study only retained available adult utterances in the transcripts, so there were many examples where a segment would be cleaned to the point of no longer containing utterances that would be analyzed. That could be the product of faulty estimations or misattribution of a speaker as an adult by LENA. In either case, there were a number of segments analyzed that did not actually contain meaningful speech for the purpose of our experiment. This impacted the statistical analysis, and required a decision on whether to retain them for the purpose of validating the sampling method. In the end, we decided to analyze only the segments that did contain adult speech.

The theoretical motivations for that decision have to do with the goal of the study and comparison to more conventional sampling. First, we do not want to put the blame of LENA's mistakes on the participants. The only reason that a sample without speech would be transcribed in the first place is because of our decision to sort and transcribe in accordance with meaningful speech. To include a segment without speech would present families as less talkative than if LENA had been more precise, but it would not necessarily indicate that a given family actually does talk less. Instead, empty segments are artifacts of LENA overestimation and our transcription pipeline. In fact, these segments are not necessarily empty at all. Depending on the individual segment, there could be a lot of child speech or other language that was excluded for not being analyzable adult speech. Regardless of the content of these segments, because we only analyzed adult speech, segments without such utterances were removed. The other motivation for

discarding those segments without adult speech is the fact that this is supposed to be a way to sample language.

In any other mode of language sampling, the continuous nature of recording would make pauses or delays contained within the sample itself. Although these empty segments are included in the sixteen hours, we are not currently observing each and every recorded segment sequentially. Instead, we are extracting a percentage of the total sample to reach an estimated word count. We are interested in the features of language across homes, and those segments neither tell us about home language, nor do they even tell us about how often a family is quiet. These segments should be discarded because they are only transcribed as the product of LENA algorithms not the home language, and if we knew ahead of time that they lacked adult speech, then they would not have been transcribed in the first place.

Future work looking to observe activity contexts in naturalistic samples with LENA could employ an activity log provided to parents. Although this would require more awareness and conscious participation in the research, it could give insight into the variation of activity contexts across families and their effects on language exchange. As previously mentioned, random segment selection would also more effectively record the natural presentation and variation of activities if an activity log is dispreferred.

This project is ongoing and the long term goal is to transcribe the full sample provided by participants for future analysis. Eventually, child language may be included in these analyses. Adult utterances were the priority of this study because ASR makes fewer errors with adult speakers, so time gains from the two-pass approach would be maximized (Shivakumar, & Georgiou, 2020). This changes the nature of our conclusions and the effects of activity contexts.

As automatic speech recognition and speaker diarization tools improve, this style of dense transcribed natural language sampling will become even more efficient..

5. Conclusion

The most important finding of this study is the validation of this sampling methodology and transcription pipeline. Results of mixed effect models and t-tests demonstrate expected variations across SES and the influence of parental input on child language performance which justifies the sampling method. Not only are LENAs able to record a denser sample of the home in a more natural way than conventional sampling, but using ASR for transcription can make that data accessible rapidly. Although topic modeling was unable to extract activity contexts, it still represents a view toward advancing techniques in the field of language sampling. As NLP tools become more sophisticated and accessible, analysis techniques should try to implement these tools as a supplement or replacement for past methodologies. In this study Whisper, the python natural language toolkit (NLTK), and gensim are all automated processes and provide output that would be human interpreted. In this way, the advanced tools are still not taken at face value and human intervention is necessary for analysis. This provides research with all of the efficiency gains associated with automation without risking making conclusions from unchecked data. The most translatable of these processes is the two-pass transcription.

By utilizing Whisper before a human trimmed the transcripts, there were time gains without any costs to accuracy. This is an entirely local process and can be operated at whatever scale. Future research could mirror this two-pass approach for the same benefits to efficiency without risks of data sensitivity. Overall, this reflects the belief that researchers have the responsibility of lowering the barrier to entry for representative data collection and analysis.

Each and every decision in terms of sampling and analysis is directly responsible for framing the results of research and clinical practice, so research must continue to improve our understanding of these decisions and their effects on conclusions drawn.

References

1. Altinkamış, N. F., Kern, S., & Sofu, H. (2014). When context matters more than language: Verb or noun in French and Turkish caregiver speech. *First Language, 34*(6), 537–550. https://doi.org/10.1177/0142723714560179
2. Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development, 92*, 484–501. https://doi.org/10.1111/cdev.13508
3. Arriaga, R. I., Fenson, L., Cronan, T., & Pethick, S. J. (1998). Scores on the MacArthur Communicative Development Inventory of children from lowand middle-income families. *Applied Psycholinguistics*, *19*(2), 209-223.
4. Baugh, J. (2017). Meaning-less differences: Exposing fallacies and flaws in “the word gap” hypothesis that conceal a dangerous “language trap” for low-income American families and their children. *International Multilingual Research Journal, 11*(1), 39–51.
5. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."
6. Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826 .
7. Bornstein, M. H., Painter, K. M., & Park, J. (2002). Naturalistic language sampling in typically developing children. *Journal of Child Language, 29*(3), 687–699.
8. Buckingham, J., Wheldall, K., & Beaman-Wheldall, R. (2013). Why poor children are more likely to become poor readers: The school years. *Australian Journal of Education, 57*(3), 190–213. https://doi.org/10.1177/0004944113495500

9. Cameron-Faulkner, T., Melville, J., & Gattis, M. (2018). Responding to nature: Natural environments improve parent-child communication. *Journal of Environmental Psychology, 59*, 9–15.
10. Crain‐Thoreson, C., Dahlin, M. P., & Powell, T. A. (2001). Parent‐child interaction in three conversational contexts: Variations in style and strategy. *New Directions for Child and Adolescent Development, 2001*(92), 23–38.
11. 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
12. 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
13. Demir-Lira, Ö. E., Applebaum, L. R., Goldin-Meadow, S., & Levine, S. C. (2019). Parents' early book reading to children: Relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental Science, 22*, e12764. https://doi.org/10.1111/desc.12764
14. Doering, E., Schluter, K., & Von Suchodoletz, A. (2020). Features of speech in German and US-American mother–toddler dyads during toy play and book-reading. *Journal of child language*, *47*(1), 112-131.

15. Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development, 92*, 609–625. https://doi.org/10.1111/cdev.13511

16. Dudley-Marling, C., & Lucas, K. (2009). Pathologizing the language and culture of poor children. *Language Arts, 86*(5), 362–370. http://www.jstor.org/stable/41483561

17. Flynn, V., & Masur, E. (2007). Characteristics of maternal verbal style: Responsiveness and directiveness in two natural contexts. *Journal of Child Language, 34*(3), 519–543. https://doi.org/10.1017/S030500090700801X

18. Gaur, Y., Lasecki, W. S., Metze, F., & Bigham, J. P. (2016, April). The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference* (pp. 1–8).

19. Gilkerson, J., Richards, J. A., & Topping, K. J. (2017). The impact of book reading in the early years on parent–child language interaction. *Journal of Early Childhood Literacy, 17*(1), 92–110. https://doi.org/10.1177/1468798415608907

20. Glas, L., Rossi, C., Hamdi-Sultan, R., Batailler, C., & Bellemmouche, H. (2018). Activity types and child-directed speech: A comparison between French, Tunisian Arabic and English. *Canadian Journal of Linguistics, 63*(4), 633–666. https://doi.org/10.1017/cnj.2018.20

21. Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly, 32*(2), 83–92. https://doi.org/10.1177/1525740110367826

22. Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics, 59*(5), 1343–1356. https://doi.org/10.1515/ling-2021-0094

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

24. Henry, D. A., Betancur Cortés, L., & Votruba-Drzal, E. (2020). Black–White achievement gaps differ by family socioeconomic status from early childhood through early adolescence. *Journal of Educational Psychology, 112*(8), 1471–1489. https://doi.org/10.1037/edu0000439

25. Hess, C. W., Haug, H. T., & Landry, R. G. (1989). The reliability of type-token ratios for the oral language of school-age children. *Journal of Speech, Language, and Hearing Research, 32*(3), 536–540. https://doi.org/10.1044/jshr.3203.536

26. Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

27. Hoff-Ginsberg, E. (1991). Mother-child conversation in different social classes and communicative settings. *Child Development, 62*(4), 782–796. https://doi.org/10.2307/1131177

28. Holme, C., Harding, S., Roulstone, S., Lucas, P. J., & Wren, Y. (2021). Mapping the literature on parent-child language across activity contexts: A scoping review. *International Journal of Early Years Education, 30*(1), 6–24. https://doi.org/10.1080/09669760.2021.2002135

29. Hsu, N., Hadley, P. A., & Rispoli, M. (2017). Diversity matters: Parent input predicts toddler verb production. *Journal of Child Language, 44*, 63–86. https://doi.org/10.1017/S0305000915000693

30. Huang, Y. T., Byrd, A. S., Asmah, R., & Domanski, S. (2023). Evaluating "meaningful differences" in learning and communication across SES backgrounds. *Annual Review of Linguistics, 9*(1), 589–608. https://doi.org/10.1146/annurev-linguistics-030521-045816

31. Hurt, H., & Betancourt, L. (2016). Effect of socioeconomic status disparity on child language and neural outcome: How early is early? *Pediatric Research, 79*, 148–158. https://doi.org/10.1038/pr.2015.202

32. Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive psychology*, *61*(4), 343-365.

33. Introducing Whisper. (2022, September 21). *OpenAI*. https://openai.com/index/whisper

34. Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology, 98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

35. Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems, 7*(24). https://doi.org/10.4108/eai.13-7-2018.163512

36. Kuchirko, Y. (2019). On differences and deficits: A critique of the theoretical and methodological underpinnings of the word gap. *Journal of Early Childhood Literacy, 19*(4), 533–562. https://doi.org/10.1177/1468798417747029

37. Lytton, H. (1971). Observation studies of parent-child interaction: A methodological review. *Child Development, 42*(3), 651–684. https://doi.org/10.2307/1127439

38. Muhinyi, A., & Hesketh, A. (2017). Low- and high-text books facilitate the same amount and quality of extratextual talk. *First Language, 37*(4), 410–427. https://doi.org/10.1177/0142723717697347

39. Murphy, K. A., Springle, A. P., Sultani, M. J., McIlraith, A., & Language and Reading Research Consortium (LARRC). (2022). Predicting language performance from narrative language samples. *Journal of Speech, Language, and Hearing Research, 65*(2), 775–784. https://doi.org/10.1044/2021_JSLHR-21-00324

40. Novotney, S., & Callison-Burch, C. (2010). Cheap, fast, and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 207–215). Association for Computational Linguistics.

41. Ogura, T., Dale, P. S., Yamashita, Y., Murase, T., & Mahieu, A. (2006). The use of nouns and verbs by Japanese children and their caregivers in book-reading and toy-playing contexts. *Journal of Child Language, 33*(1), 1–29. https://doi.org/10.1017/S0305000905007270

42. Piot, L., Havron, N., & Cristia, A. (2022). Socioeconomic status correlates with measures of Language Environment Analysis (LENA) system: a meta-analysis. *Journal of Child Language*, *49*(5), 1037–1051. doi:10.1017/S0305000921000441

43. Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*(2), 201–209. https://doi.org/10.1017/S0305000900011051

44. Rosemberg, C. R., Alam, F., Ramirez, M. L., & Ibañez, M. I. (2023). Activity contexts and child-directed speech in socioeconomically diverse Argentinian households. *International Journal of Early Childhood, 55*(1), 1–25. https://doi.org/10.1007/s13158-023-00316-w

45. Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*, 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

46. Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language, 47*(1), 5–21. https://doi.org/10.1017/S0305000919000655

47. Roy, B., Frank, M., & Roy, D. (2012). Relating activity contexts to early word learning in dense longitudinal data. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34). https://escholarship.org/uc/item/4ks237m6

48. Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(4), 264–275. https://doi.org/10.1002/wcs.1393

49. Shifrer, D., Muller, C., & Callahan, R. (2011). Disproportionality and learning disabilities: Parsing apart race, socioeconomic status, and language. *Journal of Learning Disabilities, 44*(3), 246–257. https://doi.org/10.1177/0022219410374236

50. Shivakumar, P. G., & Georgiou, P. (2020). Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations. *Computer speech & language*, *63*, 101077. https://doi.org/10.1016/j.csl.2020.101077

51. Soderstrom, M., & Wittebolle, K. (2013, November 18). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE, 8*(11), e80646. https://doi.org/10.1371/journal.pone.0080646

52. Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014, September). Automatic classification of activities in classroom discourse. *Computers & Education, 78*, 115–123. https://doi.org/10.1016/j.compedu.2014.05.010