¹   Lexical Diversity and Vocabulary Development

²   Yawen Yu[1] & Daniel Yurovsky[2]

³   [1] University of California, Los Angeles
⁴   [2] University of Chicago

⁵   Author Note

⁶   Please address corresponsdence

⁷   Correspondence concerning this article should be addressed to Yawen Yu, Postal

⁸   address. E-mail: shellyyu@uchicago.edu

Abstract

Enter abstract here. Each new line herein must be indented, like this line.

Keywords: cognitive development; language acquisition; corpus analysis

Word count: X

¹³ Lexical Diversity and Vocabulary Development

¹⁴ **Methods**

¹⁵   We report how we determined our sample size, all data exclusions (if any), all

¹⁶ manipulations, and all measures in the study.

¹⁷ **Participants**

¹⁸   This study mainly uses two large corpora—the Language Development Project (LDP)

¹⁹ and the Child Language Data Exchange System (CHILDES; MacWhinney, 2000)—to

²⁰ examine vocabulary growth of 116 typically developing children over the age of 14 months

²¹ and 58 months.

²²   The first database we use in this study is the CHILDES. It currently comprises

²³ transcripts drawn from 230 corpora in different 30 languages. This database stores a variety

²⁴ of language data, i.e. transcripts of spontaneous verbal interactions between the target child

²⁵ and their parents, siblings, and teachers, and narrative data of multilingual children, as well

²⁶ as speech samples of children with language disorders. This database is one of the most

²⁷ widely used corpus in the study of child language acquisition, particularly because of its

²⁸ sheer volume of language samples and online availability (see http://childes.psy.cmu.edu/).

²⁹   One problem, however, cautions researchers who use this database is the homogeneous

³⁰ socioeconomic status (SES) of participants. Research has documented a clear relation

³¹ between SES and child's vocabulary development (Hoff, 2003a; Rowe, 2008; Huttenlocher,

³² Vasilyeva, Waterfall, Vevea & Hedges, 2010). Given most participants in CHILDES are from

³³ high- and middle-SES, it is not clear if the language development pattern emerging from this

³⁴ corpus only represent language acquisition of children from a certain environment, and

³⁵ cannot generalize to the entire population of English-speaking children. Thus, we made use

³⁶ of another large corpus—LDP corpus— in which participants were selected to match as

³⁷ closely as possible the 2000 census data on family income and ethnicity range in the great

³⁸ Chicago area.

LDP is a longitudinal research project conducted since 2002 with its focus on linguistic and gestural inputs and their consequences for child language and gesture in both typically developing and brain injured children. This is a collection of transcripts of conversations between the target child and the parental caregiver(s) for a 90-minute period at each visit. All conversations are videotaped during ordinary daily interactions every four months for a total of 12 visits between 14 months and 58 months.

For our purpose, we primarily examined lexical development of 66 typically-developing children from a diverse demographic background. Additionally, in this project, researchers also assessed child's vocabulary skills with the use of MacArthur-Bates Communicative Development Inventories (CDIs) at 14 months and Peabody Picture Vocabulary Test (PPVT) at 30, 42 and 54 months, respectively. These two measures have been widely used as standard instruments to assess vocabulary acquisition and to diagnose specific language impairment (SLI) in children (Eickhoff, Betz, & Ristow, 2010). Given normative information of individual language development is difficult to derive from observational data because a spontaneous language sample is particularly sensitive to high-frequency words (Dale&Fenson; 1996), the CDI and PPVT would serve as a valid comparison for growth in other indicators of vocabulary acquisition. Notably, CHILDES doesn't contain normative data provided by external measurements, thus, mean length of utterances in morphemes(MLUm) is calculated in CHILDES as well as LDP corpus.

LDP and CHILDES corpora provide us with complementary data to investigate vocabulary development of children from a diverse demographic background. The criteria for drawing the sample used in the present study were the following. First, given the goal of this study is to examine lexicon development of typically-developing children, children with language impairment or brain injury were eliminated. Second, children whose home language was not English (e.g. North-American English and British English) were excluded, as the language development of bilingual children were not considered to be comparable to that of children who exclusively speak and hear English at home. Third, of the remaining 760

66 children, we only collected language samples of children whose interactions with parents were

67 videotaped for at least three sessions between 14 month and 58 months, as adequate

68 language data is required for constructing accurate individual vocabulary growth. The

69 tokens that were transcribed and counted included all dictionary words, onomatopoeic

70 sounds (e.g. da-da), and evaluative sounds (e.g. uh-oh). The final sample for the present

71 study includes 116 primary caregiver-child dyads. Both corpora contain a total of about 12

72 million tokens after removing a number of special transcription characters and other artifacts

73 of the CHILDES coding system, as well as un-transcribable sections

74 **Material**

75 **Procedure**

76 **Data analysis**

77   The present study concerns children's vocabulary growth, especially growth of lexical

78 diversity. To address this issue, we demonstrated analytically how growth curve parameters

79 change in a deterministic manner under different lexical diversity measures and how

80 variations in measures influence understanding the role of caregiver's input on language

81 outcome of children. It is difficult to establish the role of input, because of two nagging third

82 variable-problem: (1) Shared variability in linguistic diversity between parents and children

83 reflects context rather than process, and (2) That variability in both input and output are

84 explained by a common variable (e.g. some non-environmental genetic variable).

85   We tackled both of these problems by using growth-curve analyses that allow us to

86 separate each participant's intercept—a measure that captures individual initial

87 aptitude—from their rate of development. We apply this analysis to both child and caregiver

88 speech, in order to determine which aspects of development differ across children and which

89 aspects of input may influence development. We employed mixed-effect model to construct a

90 growth trajectory for each participant over an extended time period from 14 to 58 months.

91   Trajectories of children's vocabulary development are described by two person-specific

₉₂ parameters: intercept and slope. Mixed-effects models allow us to consider all factors that

₉₃ potentially contribute to the growth of children's vocabulary. These factors comprise not

₉₄ only standard fixed-effects factors, more specifically, average expected lexical diversity value

₉₅ across children and across sessions, but also covariates bound to the subjects.

₉₆   Another advantage of mixed-effects model is that local dependencies between the

₉₇ successive measures, specifically, vocabulary skills in preceding sessions, can be brought into

₉₈ the model. Lastly, it is particularly useful for handling situations in which measures for some

₉₉ individuals are missing at some time point. Overall, mixed-effects models allow for the

₁₀₀ subject and age specific adjustments to intercept and slope, and thus, enhanced precision in

₁₀₁ prediction and estimation. Given measured lexical diversity changes as a function of the log

₁₀₂ age, slope in the present study is characterized as linear growth in a form of log age, and

₁₀₃ intercept is defined as expected individual lexical diversity value at 30 months.

₁₀₄   After constructing individual growth trajectories, we turn to three fundamental

₁₀₅ questions in order to address the primary concern of this paper. The first question is

₁₀₆ whether the overall trajectories of children and caregivers speech change over time.

₁₀₇   The second question is whether there are significant individual differences among

₁₀₈ participants in LDP and CHILDES corpora. We used mixed-effects models to investigate

₁₀₉ variations in emphasized growth curve parameters with respect to different lexical diversity

₁₁₀ indices (e.g. MTLD, TTR, vocd-D and MATTR). Therefore, we tracked not only the overall

₁₁₁ characteristics of participants' vocabulary development, but also the nature of individual

₁₁₂ differences in their pattern of language use.

₁₁₃   If there are significant variations in child's growth parameters, the third question is

₁₁₄ what factors can predict child's vocabulary growth across time. Here, we evaluate possible

₁₁₅ correlations among the components of child's and caregiver's vocabulary growth. Abundant

₁₁₆ research has demonstrated associations between maternal language and child's early lexicon

₁₁₇ development (e.g., Hart & Risley, 1995; Hoff, 2003; Huttenlocher, Haight, Bryk, Seltzer &

₁₁₈ Lyons, 1991; Huttenlocher, Waterfall, Vasilyeva, Vevea & Hedges, 2010; Pan, Rowe, Singer &

Snow, 2005; Rowe, 2008). However, it remains unknown whether these correlations vary with different indices used to measure vocabulary skills. We compared the parameters deriving from lexical diversity indices of our primary interest to that of normative measures, including PPVT, CDI and the mean length of utterance(MLU).

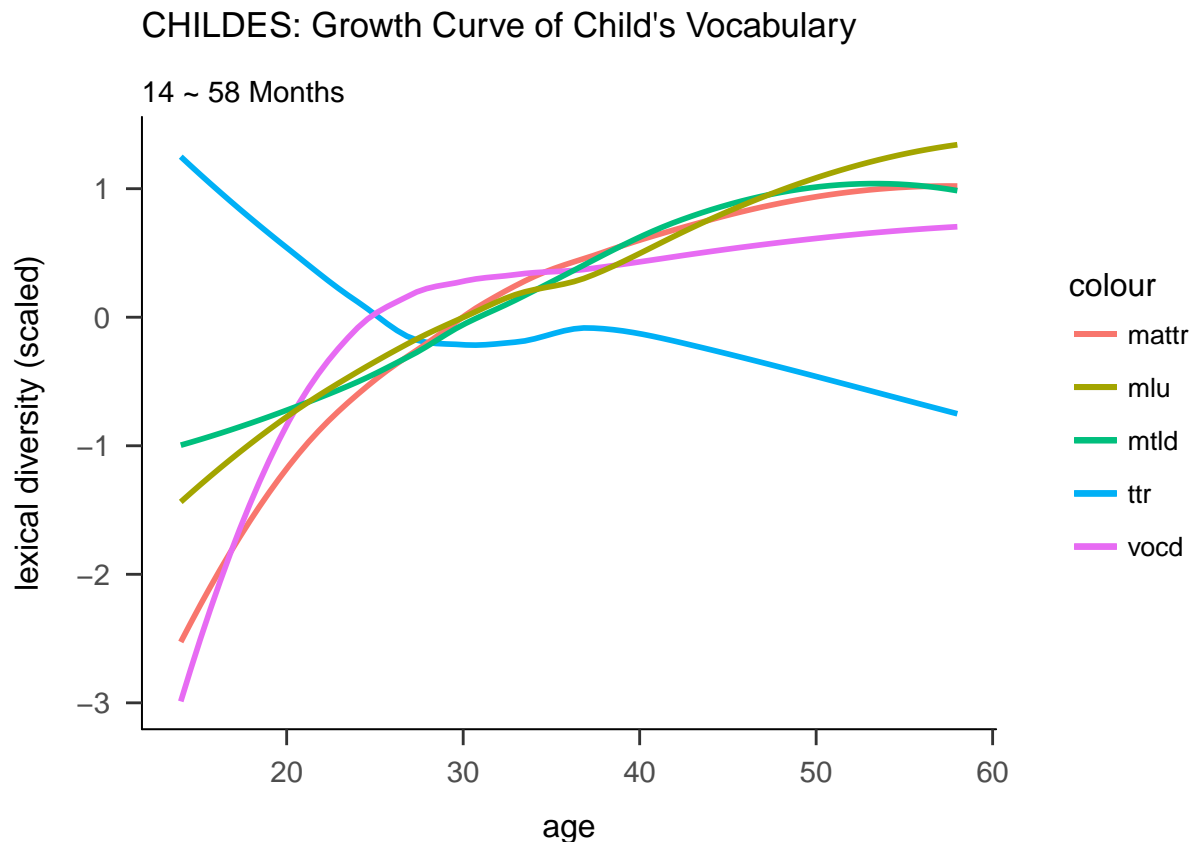*Figure 1*

Results

Growth curve of child's vocabulary

The first goal of the study is to examine whether lexical diversity measures of children change over time. We plot growth trajectories of child's vocabulary skills measured by different methods at each session during 2;2 and 4;10. All measures are scaled based on their standard deviation and mean, thus, could be presented in one figure. Figure 1 shows accelerating curves of MTLD, MATTR, vocd-D, that are characterized by a log-linear shape.

130 These curves resemble the curves of MLU in both corpora and PPVT, CDI vocabulary as

131 well as sentence complexity in LDP. All measures, except for TTR curve, increase from 14 to

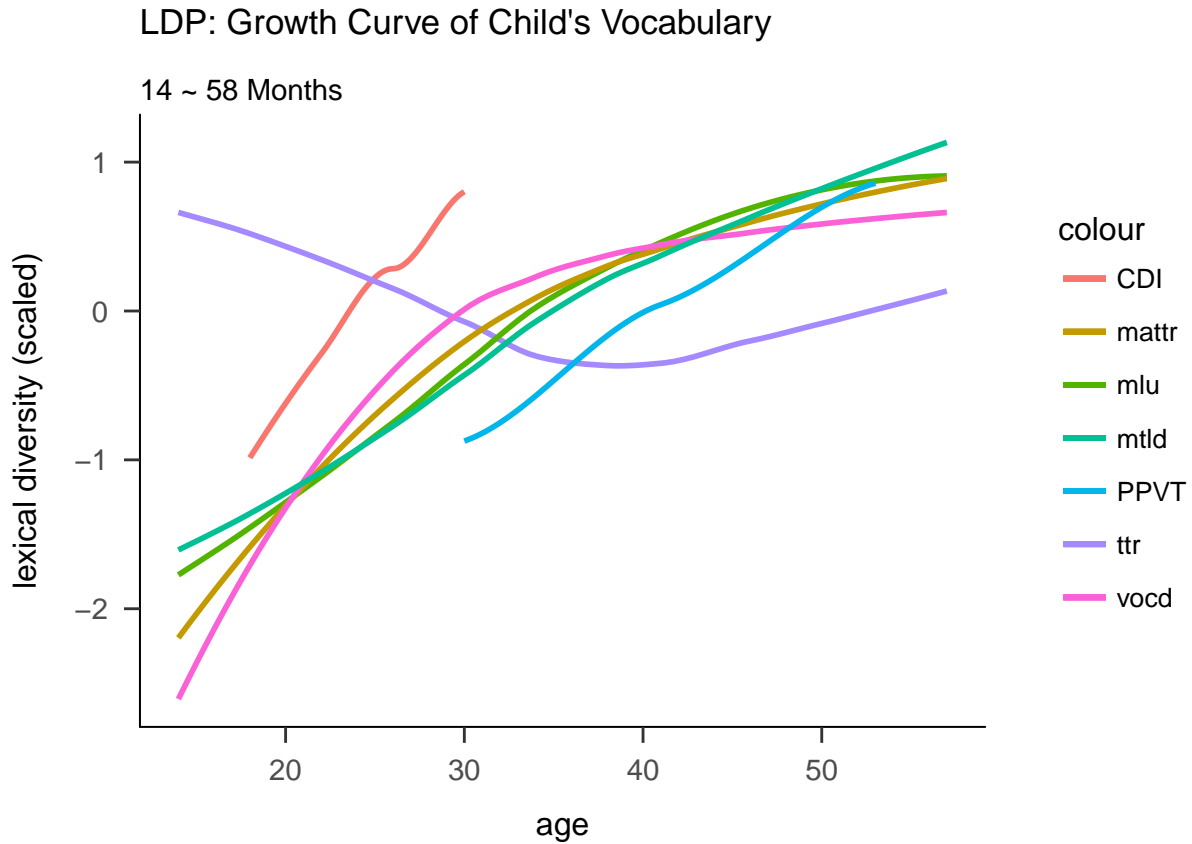132 58 months and growth gradually diminishes over time for vocd-D.

Figure 2

133   We further fit regression models to evaluate relation between child's intercept and age.

134 As expected, child's initial status of vocabulary skills are significantly related to age. For

135 example, in LDP corpus, age is a strong predictor of the intercepts deriving from MTLD

136 (r=0.85, p<0.001), MATTR (r=0.82, p<0.001), vocd-D (r=0.71, p<0.001), that is similar to

137 the normative measures: CDI (r=0.93, p<0.001), PPVT (r=0.95, p<0.001) and MLU

138 (r=0.84, p<0.001). By comparison, age explains less variance in TTR measures (r=0.46,

139 p<0.01). TTR curve is the most volatile and cannot represent the growth pattern of child's

140 lexicon over time. So far, the results concur with findings in many previous research (Heaps,

141 1978; Herdan, 1960; Arnaud, 1984; Kucera & Francis, 1967; Montag, Jones, & Smith, 2018)

that TTR, also known as type-token ratio, demonstrates diminishing returns of new types. Therefore, when it is used to compare any two texts, the longer one generally appears to be less diverse.

Variation in vocabulary development

The second goal is to document individual differences in child's vocabulary development and caregiver's child-directed speech. We first fit all vocabulary measures, assessed by MTLD, MATTR, vocd-D, TTR, MLU, PPVT and CDI vocabulary and sentence complexity (the last three measures are only available in LDP corpus), with log of age as a sole predictor. We obtain paramters of growth trajectories, specifically, the intercept describing vocabulary skills at 2;6 and the slope demonstrating the change rate over time. Descriptive statistics for these parameters are presented in Table I. Coefficient of variation is computed by dividing mean of each measure by their standard deviation. Results show that children varied widely in the initial vocabulary skills and the rate of lexical development, however, variance siginificantly differ with respect to various measures. For example, there is the largest variation in the slope measured by type-token ratio, which is approximately 3 times as the child's slope drew from vocd-D. At the level of intercept, a large amount of variance in the intercept assessed by CDI sentence complexity is noticeable in child's speech, that is over 8 times as the intercept deriving from vocd-D. The third goal of the study is to evaluate predictors of growth parameters of child's lexical development.
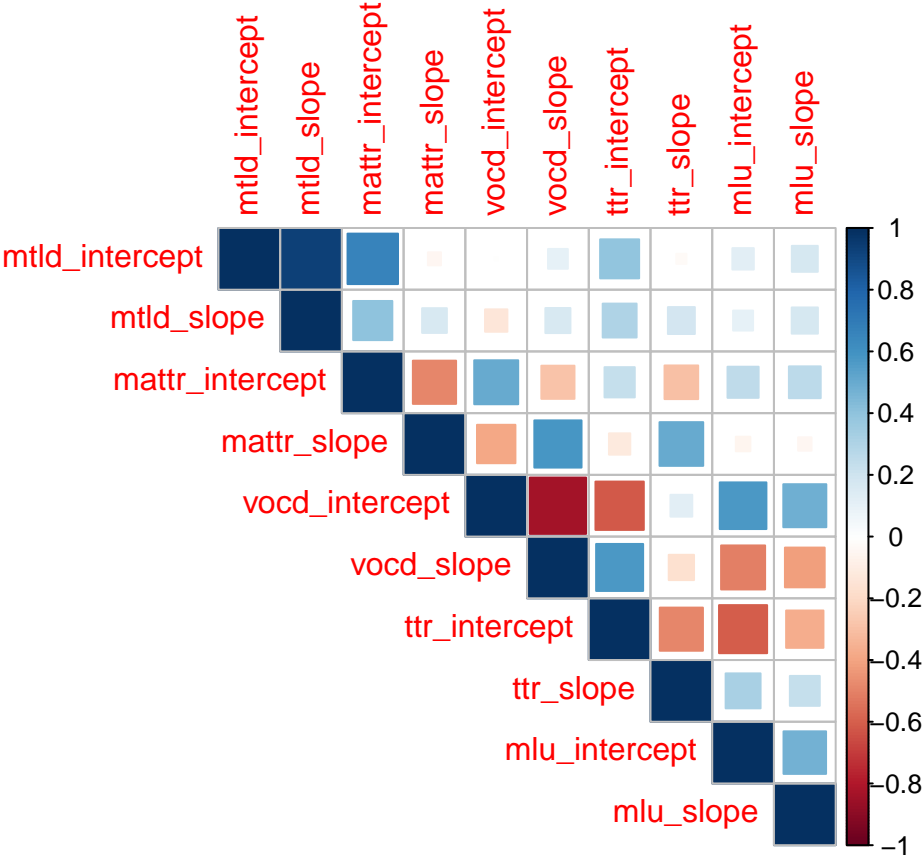
Correlation

Children vary widely in their intercept and slope of vocabulary growth trajectory. Results show significant relation between CHILDES child's intercept and slope, but relations vary with different measures. For example, there are significant positive relations between child's intercept deriving from MTLD and the slope from MTLD in both corpora. But the intercept drawing from MATTR, vocd-D and TTR negatively relate to the slope from these measures, respectively. Given the homogenous background of CHILDES corpus, we conduct

Table 1

Descriptive statistics for child's speech measures at child age 2;6 and growth rate (n=116)

measure	mean	sd	cv
cdi_intercept	497.71	151.16	0.30
cdi_slope	834.32	130.35	0.16
mattr_intercept	0.48	0.08	0.17
mattr_slope	0.22	0.08	0.34
mlu_intercept	2.88	0.78	0.27
mlu_slope	1.98	0.69	0.35
mtld_intercept	14.24	3.94	0.28
mtld_slope	18.85	7.82	0.42
ppvt_intercept	27.45	10.91	0.40
ppvt_slope	77.49	18.15	0.23
sen_intercept	19.26	11.29	0.59
sen_slope	39.00	19.08	0.49
ttr_intercept	0.23	0.10	0.42
ttr_slope	-0.06	0.09	-1.56
vocd_intercept	29.58	2.17	0.07
vocd_slope	10.30	5.56	0.54

the same correlation analysis on the more diverse LDP corpus. Fugure 2 demonstrates the same results as that in CHILDES. Because the results of correlations from different measures are in contrast to each other, we compare the correlations from target measures to that emerging from normative data, in particularly, CDI vocabulary, CDI sentence complexity, PPVT and MLU. Results show there are significant positive relations between child's intercept and slope deriving from nomative measures, that concur with MTLD. In sum, children who produced a more diverse lexicon at 2;6 grow in a faster rate in terms of vocabulary diversity. MTLD, among other measure, best depicts this correlation.

176

177
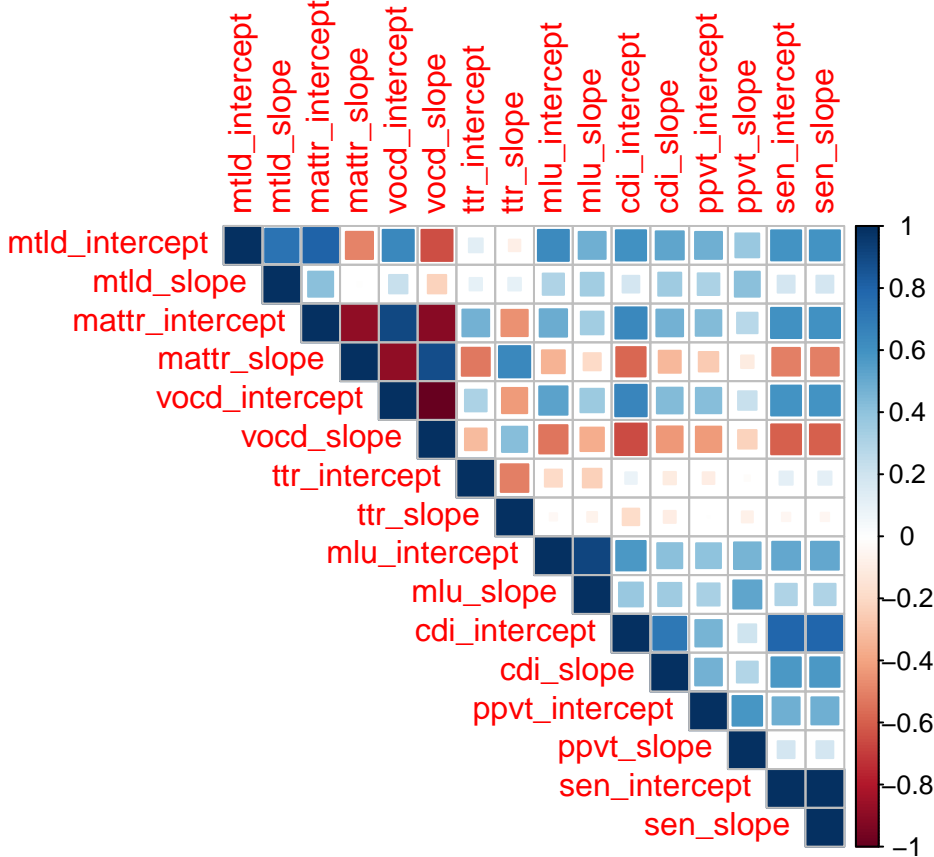
178   Furthermore, a body of previous work shows significant influence of caregiver's speech

179 on child's language development (Rowe, 2008;...). Figure 3 demonstrates such a relation

180 among gwoth parameters drawn from MTLD of paticipants in CHILDES, but not other

181 three measures. Specifically, caregiver's intercept deriving from MTLD and TTR are

182 significant positive predictors of child's intercept deriving from the same measures, while

183 such correlations are negligible in MATTR and vocd-D. As is discussed earlier, such a

184 correlation between lexical diversity of caregiver and child at 30 months relates not only to

185 genetic reasons, but also to conversational contexts. Paticipants who read picture books tend

186 to yield more diverse vocabulary than those talk during meal time(...). On the other hand,

187 the significant relation between caregiver's intercept and child's slope is only shown in

188 MTLD measure. The same results are also found in LDP corpus, in which correlations

189 between caregiver's speech and child's speech are siginificant in MTLD measure. When

190 comparing to normative data, results generated by MTLD concur with MLU measure.

191
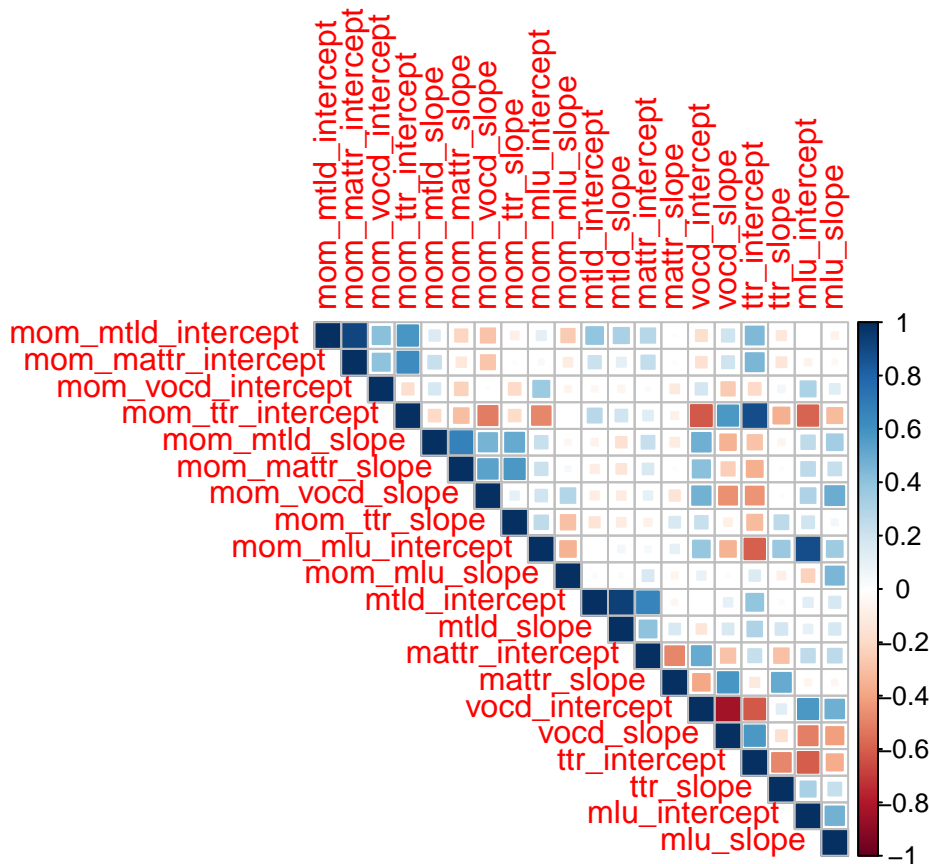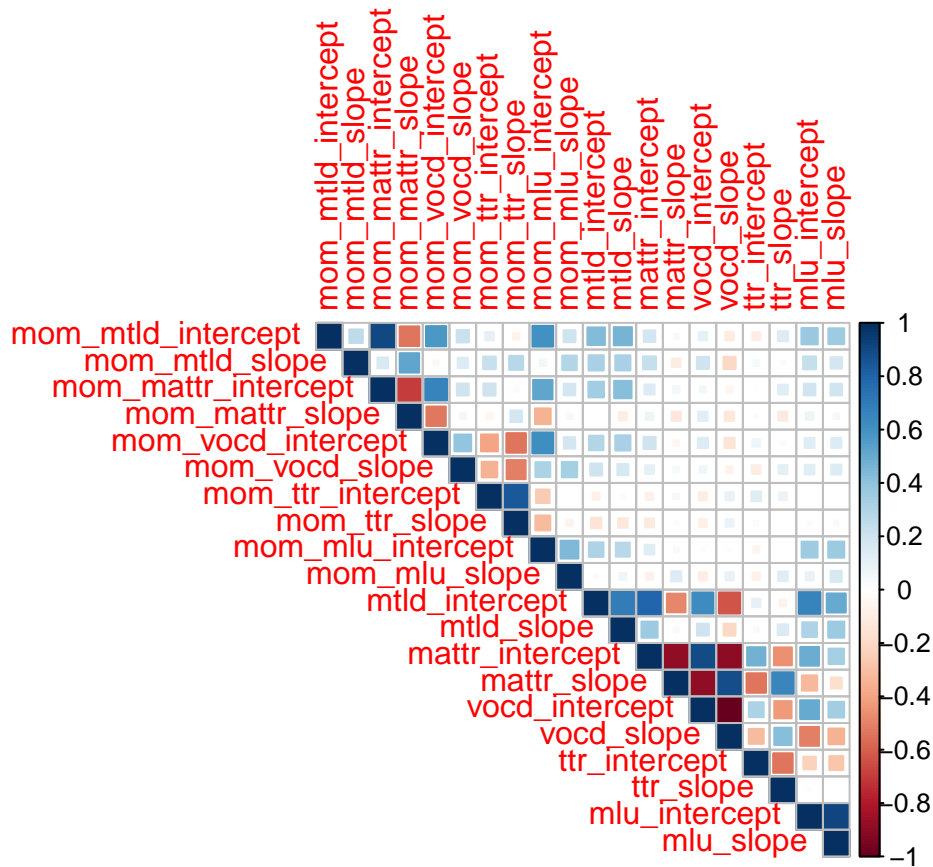
The mechanism of MTLD

The correlation analysis demonstrates that caregiver's lexical diversity at early age is a significant predictor of child's initial vocabulary skill, and child's initial vocabulary skill and early speech of caregiver are strong predictors of the growth rate of child's vocabulary, as measured by MTLD and normative measures. So far, the results generated by MTLD concur with previous findings. To explore what distinguishes MTLD from other lexical diversity techniques (i.e. vocd-D, TTR and MATTR), we examine its theoretical rationale and test how this mechanism works using simulation.

Sequential analysis

Conceptually, MTLD estimates average number of consecutive tokens for which a certain TTR is maintained (e.g. 0.72 by default). For any given sample, each token is

₂₀₄ evaluated sequentially for its TTR. For example, "I"(TTR = 1) "had"(TTR = 1)

₂₀₅ "chicken"(TTR = 1) "and" (TTR = 1) "I" (TTR = 0.8) "also" (TTR = 0.83) "had" (TTR =

₂₀₆ 0.71) and so forth. When the default TTR score is reached (here, 0.72), the factor count

₂₀₇ increases by a value of 1 and the TTR evaluations are reset. This process is repeated until

₂₀₈ the last token of the sample is evaluated for its TTR. Then the total number of tokens is

₂₀₉ divided by the total factor count. Subsequently, the same process is repeated on the reversed

₂₁₀ language sample. The final MTLD value is the mean of forward and reversed MTLD scores.

₂₁₁   When looking into existing lexical diversity indices, nonsequential analysis is still a

₂₁₂ common approach. One reason of its being ubiquitous relates to the advantage of avoiding

₂₁₃ local clustering. However, it may lead to a distorted way of overall text (Malvern et al.

₂₁₄ 2004). MTLD is an exception. The sequential analysis of MTLD distinguishes itself from

₂₁₅ other measures by maintaining the integrity of a text, because it evaluates words in order,

₂₁₆ rather than treats a text as a bag of words. Words, or ther textual components, have to be

₂₁₇ bound together with a certain structure so that a reader or a listener can form a coherent

₂₁₈ mental representation (Van Dijk & Kintsch, 1983). Therefore, the sequential analysis may

₂₁₉ provide information on vocabulary from various levels, lexical level and sematic level, that

₂₂₀ interact in an intricate way. The final set of analyses explore how MTLD works differently

₂₂₁ from other measures by assessing multiple simulated text sampled from LDP.

₂₂₂ **Simulated speech**

₂₂₃   The sequential analysis differs from nonsequential analysis mainly in its measuring a

₂₂₄ text in order. We first evaluate whether the order of words in a text influences lexical

₂₂₅ diversity score of the entire text. To test this, we begin with a baseline sample of 3000 tokens

₂₂₆ from large LDP corpus and then simulate another two simulated child speech generated by

₂₂₇ including 15 tokens in a repetitive order or in a random order. For the 15 tokens, we

₂₂₈ generate a list of all the unique word types produced by children in the entire corpus, and

₂₂₉ select the first 5 word types that occur in LDP most frequently, specifically, "I", "you", "the",

"it" and "no". In the second sample, we add a total number of 15 tokens with each word type

repeating 3 times in such a repetitive order as "i", "i", "i", "no", "no", "no", "you", "you",

"you", "the", "the", "the", "it", "it", "it". The third sample is created by inserting the same

5 word types in a random order. We then repeat this sampling procedure 100 times and

measure three types of child speech with four lexical diversity techniques. Results are shown

in Figure 4. Whereas vocd-D and TTR scores decrease when add 15 tokens regardless of

their orders, there is a consistent decrease in MTLD scores when comparing samples of

various word orders, though the order is manipulated at a very small scale of 0.5 percent

tokens. However, it remains unclear whether the decrease in MTLD scores is caused by

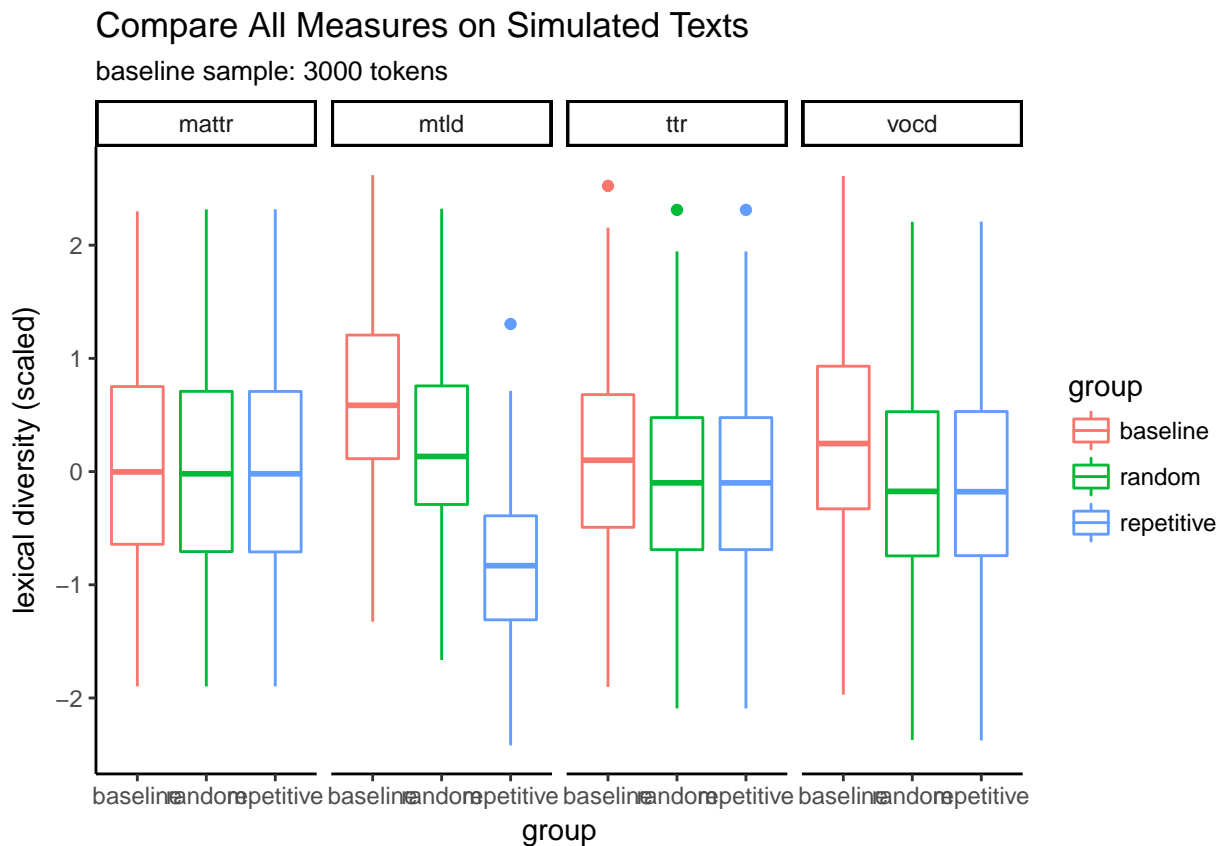different word orders or the high frequency words that actually yield more repetitions.

Figure 3

The second question emerging from this is whether word frequency influences lexical

diversity score and whether the effect varies with respect to different measures. We also

242 randomly sample 3000 tokens as a basline child speech and add 5 unique low-frequency word

243 types in a repetitive order or in a random order, respectively. To be more specific, these

244 word types are "treatment", "clog", "trustworthy", "thief" and "tofu"; each word type only

245 occurs once in the LDP corpus. The second sample comprises of the baseline sample with

246 these 5 unique word types repeating 3 times in order, and the third sample entails these 5

247 word types repeating 3 times in a random order. We perform the same sampling procedure

248 described previously 100 times. Figure 5 demonstrates MTLD scores are sensitive to word

249 order regardless of the frequency of words added. Whereas MATTR and TTR are influenced

250 by neither word frequency or word order, vocd-D scores slightly increase when low-frequency

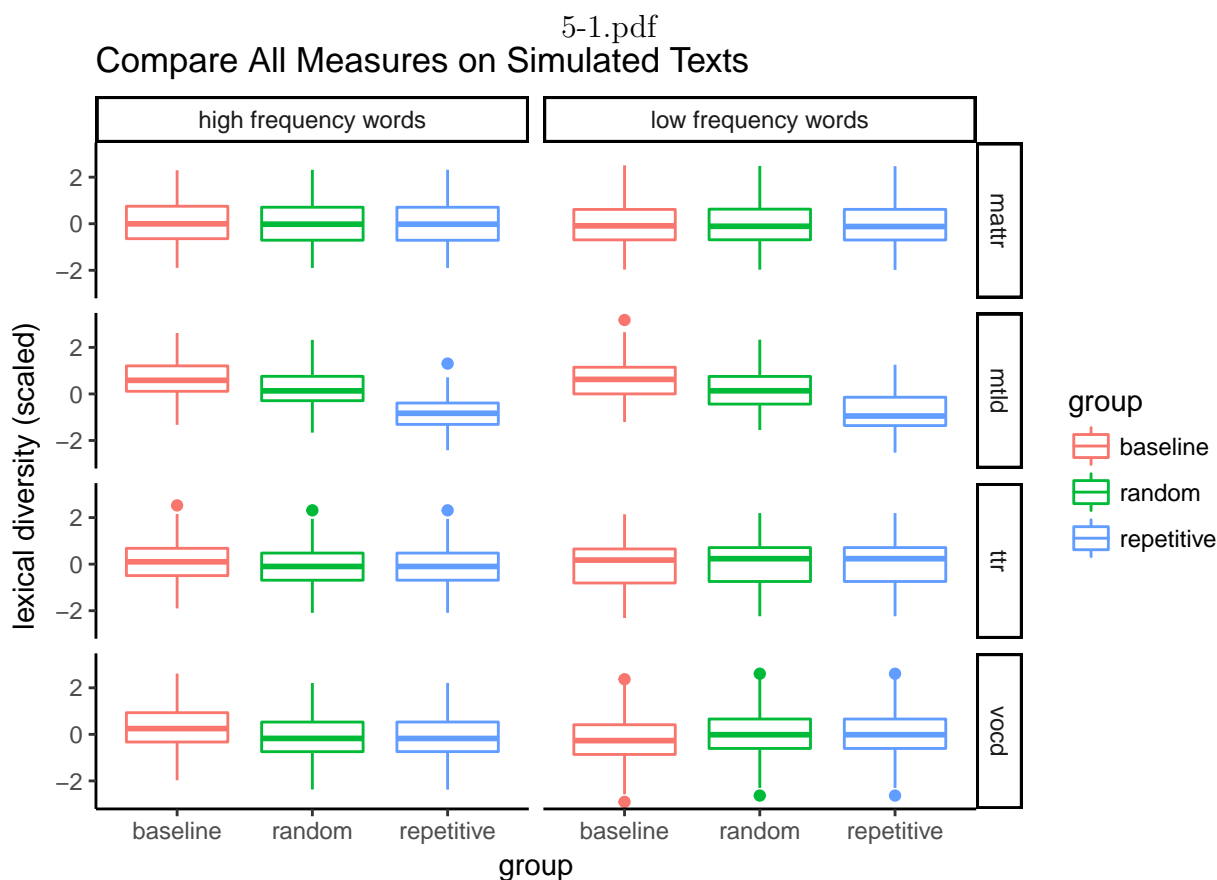251 words are added but decrease when adding high-frequency words.

Figure 4

252 <div align="center">**Discussion**</div>

253 <div align="center">**Acknowledgements**</div>

257 # References