Lexical Diversity and Vocabulary Development

Yawen Yu[1] & Daniel Yurovsky[2]

[1] University of California, Los Angeles

[2] University of Chicago

Author Note

Please address corresponsdence

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

Lexical Diversity and Vocabulary Development

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

This study mainly uses two large corpora—the Language Development Project (LDP) and the Child Language Data Exchange System (CHILDES; MacWhinney, 2000)—to examine vocabulary growth of 150 typically developing children over the age of 14 months and 58 months.

The first database we use in this study is the CHILDES. It currently comprises transcripts drawn from 230 corpora in different 30 languages. This database stores a variety of language data, i.e. transcripts of spontaneous verbal interactions between the target child and their parents, siblings, and teachers, and narrative data of multilingual children, as well as speech samples of children with language disorders. This database is one of the most widely used corpus in the study of child language acquisition, particularly because of its sheer volume of language samples and online availability (see http://childes.psy.cmu.edu/).

One problem, however, cautions researchers who use this database is the homogeneous socioeconomic status (SES) of participants. Research has documented a clear relation between SES and child's vocabulary development (Hoff, 2003a; Rowe, 2008; Huttenlocher, Vasilyeva, Waterfall, Vevea & Hedges, 2010). Given most participants in CHILDES are from high- and middle-SES, it is not clear if the language development pattern emerging from this corpus only represent language acquisition of children from a certain environment, and cannot generalize to the entire population of English-speaking children. Thus, we made use of another large corpus—LDP corpus— in which participants were selected to match as closely as possible the 2000 census data on family income and ethnicity range in the great 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.

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 children, we only collected language samples of children whose interactions with parents were videotaped for at least five sessions between 14 month and 58 months, as adequate language

data is required for constructing accurate individual vocabulary growth. The tokens that were transcribed and counted included all dictionary words, onomatopoeic sounds (e.g. da-da), and evaluative sounds (e.g. uh-oh). The final sample for the present study includes 150 primary caregiver-child dyads. Both corpora contain a total of about 15 million tokens after removing a number of special transcription characters and other artifacts of the CHILDES coding system, as well as un-transcribable sections

Material

Procedure

Data analysis

The present study concerns children's vocabulary growth, especially growth of lexical diversity. To address this issue, we demonstrated analytically how growth curve parameters change in a deterministic manner under different lexical diversity measures and how variations in measures influence understanding the role of caregiver's input on language outcome of children. It is difficult to establish the role of input, because of two nagging third variable-problem: (1) Shared variability in linguistic diversity between parents and children reflects context rather than process, and (2) That variability in both input and output are explained by a common variable (e.g. some non-environmental genetic variable).

We tackled both of these problems by using growth-curve analyses that allow us to separate each participant's intercept—a measure that captures individual initial aptitude—from their rate of development. We apply this analysis to both child and caregiver speech, in order to determine which aspects of development differ across children and which aspects of input may influence development. We employed mixed-effect model to construct a growth trajectory for each participant over an extended time period from 14 to 58 months.

Trajectories of children's vocabulary development are described by two person-specific parameters: intercept and slope. Mixed-effects models allow us to consider all factors that potentially contribute to the growth of children's vocabulary. These factors comprise not

only standard fixed-effects factors, more specifically, average expected lexical diversity value across children and across sessions, but also covariates bound to the subjects.

Another advantage of mixed-effects model is that local dependencies between the successive measures, specifically, vocabulary skills in preceding sessions, can be brought into the model. Lastly, it is particularly useful for handling situations in which measures for some individuals are missing at some time point. Overall, mixed-effects models allow for the subject and age specific adjustments to intercept and slope, and thus, enhanced precision in prediction and estimation. Given measured lexical diversity changes as a function of log transformed age, slope in the present study is characterized as linear growth in a form of log age, and intercept is defined as expected individual lexical diversity value at 30 months.

After constructing individual growth trajectories, we turn to three fundamental questions in order to address the primary concern of this paper. The first question is whether the overall trajectories of children and caregivers speech change over time.

The second question is whether there are significant individual differences among participants in LDP and CHILDES corpora. We used mixed-effects models to investigate variations in emphasized growth curve parameters with respect to different lexical diversity indices (e.g. MTLD, TTR, vocd-D and MATTR). Therefore, we tracked not only the overall characteristics of participants' vocabulary development, but also the nature of individual differences in their pattern of language use.

If there are significant variations in child's growth parameters, the third question is what factors can predict child's vocabulary growth across time. Here, we evaluate possible correlations among the components of child's and caregiver's vocabulary growth. Abundant research has demonstrated associations between maternal language and child's early lexicon development (e.g., Hart & Risley, 1995; Hoff, 2003; Huttenlocher, Haight, Bryk, Seltzer & 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 generated

by lexical diversity indices of our primary interest to that of normative measures, including PPVT, CDI vocabulary and CDI sentence complexity measures.

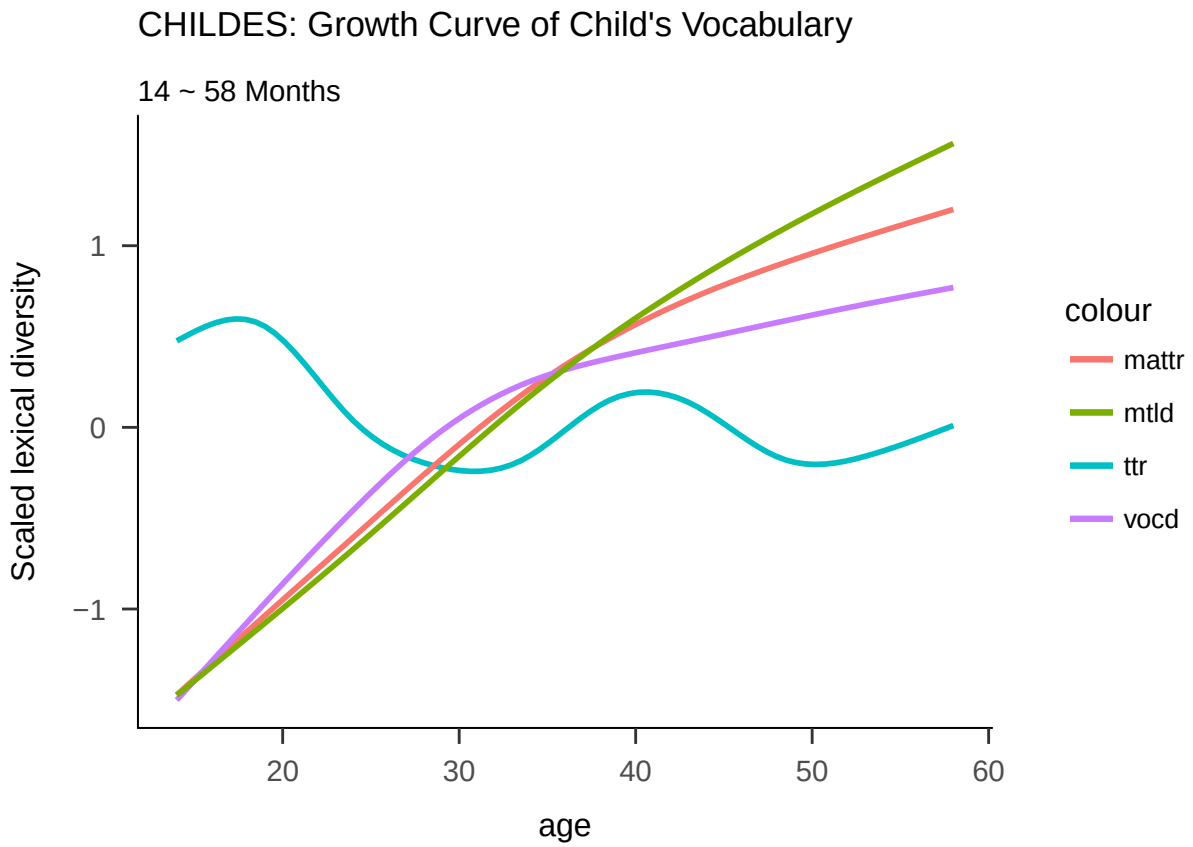


Figure 1

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 presents accelerating curves of children's vocabulary growth in CHILDES generated by MTLD, MATTR and vocd-D, that are characterized by a log-linear shape. We also plot the curves of PPVT, CDI vocabulary and sentence complexity measures as external norms. The curves in

LDP corpus display the similar log-linear shape, showin in Figure 2. All measures, except for TTR curve, increase from 14 to 58 months and growth gradually diminishes over time for vocd-D.

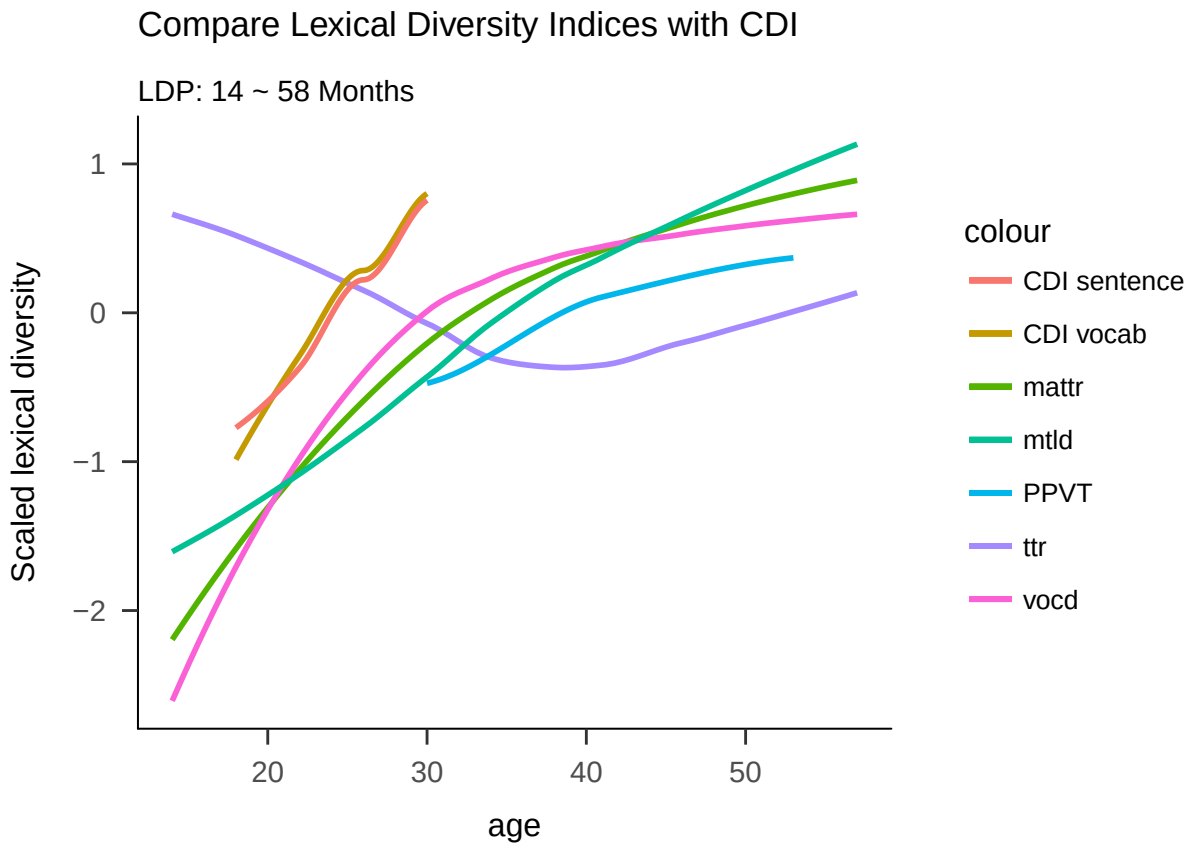


Figure 2

We further fit regression models to evaluate relation between child's intercept and age. As expected, child's initial status of vocabulary skills are significantly related to age. For example, in LDP corpus, age is a strong predictor of the intercepts deriving from MTLD ($r=0.85$, $p<0.001$), MATTR ($r=0.82$, $p<0.001$), vocd-D ($r=0.71$, $p<0.001$), that is similar to the normative measures: CDI ($r=0.93$, $p<0.001$) and PPVT ($r=0.95$, $p<0.001$). By comparison, age explains less variance in TTR measures ($r=0.46$, $p<0.01$). TTR curve is the most volatile and cannot represent the growth pattern of child's lexicon over time. So far, the results concur with findings in many previous research (Heaps, 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, PPVT and CDI vocabulary and sentence complexity, with log transformed age as a sole predictor. The last three measures are only available to 62 children in LDP corpus. We obtain paramters of growth trajectories, specifically, the intercept describing vocabulary skills at 2;6 and the slope showing the rate of vocabulary development 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 display that children don't varied widely in the initial vocabulary skills, but in the rate at which children develop lexicon richness and the variance siginificantly differ with respect to various measures. For example, in CHILDES corpus, the largest variation in the slope is measured by type-token ratio, that is approximately 9 times as the child's slope drew from MTLD. The analysis on variations of child's growth parameters in a more diverse LDP sample revealed parallel results, shown in Table 2. Comparing to CHILDES corpus, the variance in child's slope is much smaller. The third goal of the study is to evaluate predictors of growth parameters of child's lexical development.

Correlation

Children vary widely in their slope of vocabulary growth trajectory. We first compare growth parameters generated by MTLD, MATTR, vocd-D and TTR to normative measures in the LDP sample. Results shows a significant correlation between child's intercept generated by MTLD, MATTR, vocd-D and that by CDI and PPVT, while child's slope only deriving from MTLD highly correlates to all normative measures.

Table 1

Descriptive statistics for CHILDES child's speech measures at child age 2;6 and growth rate (n=84)

measure	mean	sd	cv
mattr_intercept	0.49	0.04	0.09
mattr_slope	0.18	0.09	0.54
mtld_intercept	17.31	4.56	0.26
mtld_slope	21.10	4.25	0.20
ttr_intercept	0.23	0.10	0.42
ttr_slope	-0.05	0.09	-1.86
vocd_intercept	32.03	1.77	0.06
vocd_slope	6.16	5.53	0.90

We then evaluate predictors of the growth rate of child's vocabulary in both corpora. Results present significant negative relations between child's intercept and slope in CHILDES dataset. Given the homogenous background of CHILDES corpus, we conduct the same correlation analysis on the more diverse LDP corpus. Fugure 2 shows a significant negative correlation between intercept and slope of children in LDP, generated by MATTR, vocd-D and TTR, but a significant positive relation deriving from MTLD. We compare the correlations from these measures to normative data. Child's intercept positively relates to their slope deriving from nomative measures, that is consistent with MTLD.

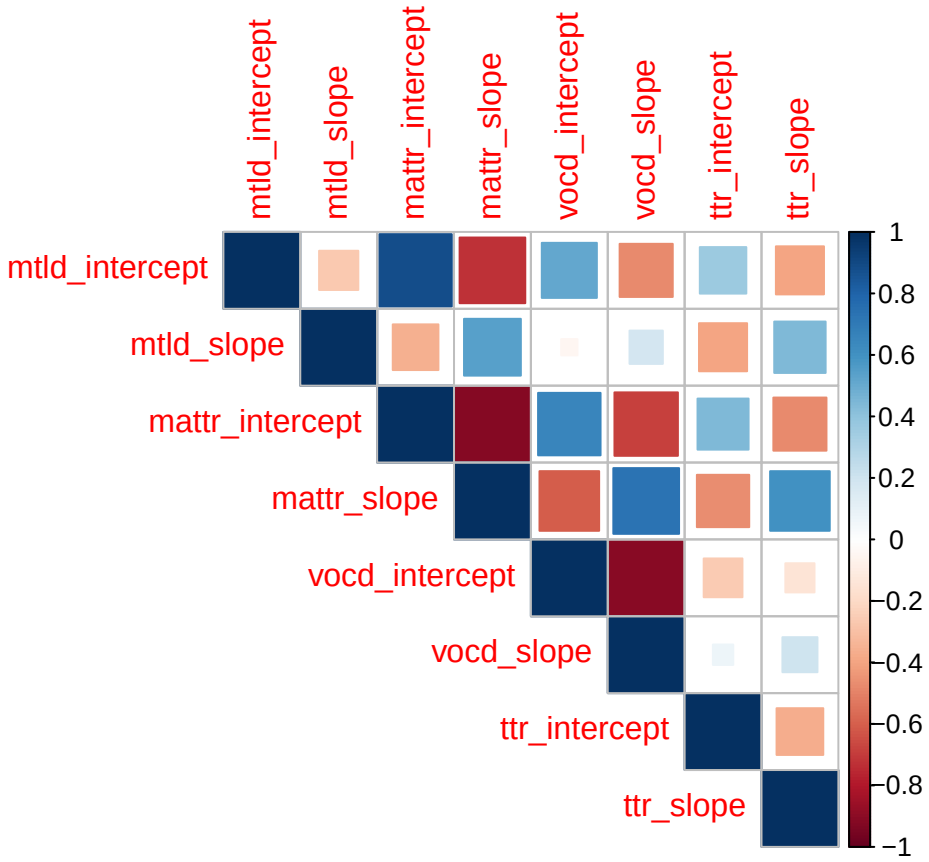
```
## NULL
```

```
## NULL
```

Table 2

Descriptive statistics for LDP child's speech measures at child age 2;6 and growth rate (n=62)

measure	mean	sd	cv
cdi_intercept	497.71	151.16	0.30
cdi_slope	834.32	130.35	0.16
mattr_intercept	0.42	0.04	0.10
mattr_slope	0.23	0.05	0.22
mtld_intercept	12.57	2.31	0.18
mtld_slope	17.94	3.01	0.17
ppvt_intercept	43.39	20.73	0.48
ppvt_slope	50.69	16.79	0.33
sen_intercept	19.26	11.29	0.59
sen_slope	39.00	19.08	0.49
ttr_intercept	0.19	0.03	0.14
ttr_slope	-0.03	0.05	-1.57
vocd_intercept	29.22	1.75	0.06
vocd_slope	10.28	2.88	0.28



3-1.pdf

Figure 3

Furthermore, a body of previous work demonstrates significant influence of caregiver's speech on child's language development (Rowe, 2008;. . .). Figure 3 demonstrates that such a relation among gwoth parameters occurs in MTLD measure, but not other three measures. Specifically, caregiver's intercept deriving from MTLD, MATTR and TTR are significant positive predictors of child's intercept generated by the same measures, while the correlation is negligible in vocd-D. As is discussed earlier, the correlation between lexical diversity of caregiver and child at 30 months relates not only to genetic reasons, but also to conversational contexts. Paticipants who read picture books tend to produce more diverse vocabulary than those talk during meal time(. . .). On the other hand, caregiver's slope has a marginally significant relation to child's slope generated by MTLD measure. Using the same analytical approach as performed for CHILDES, seperated correlation analysis tests relation

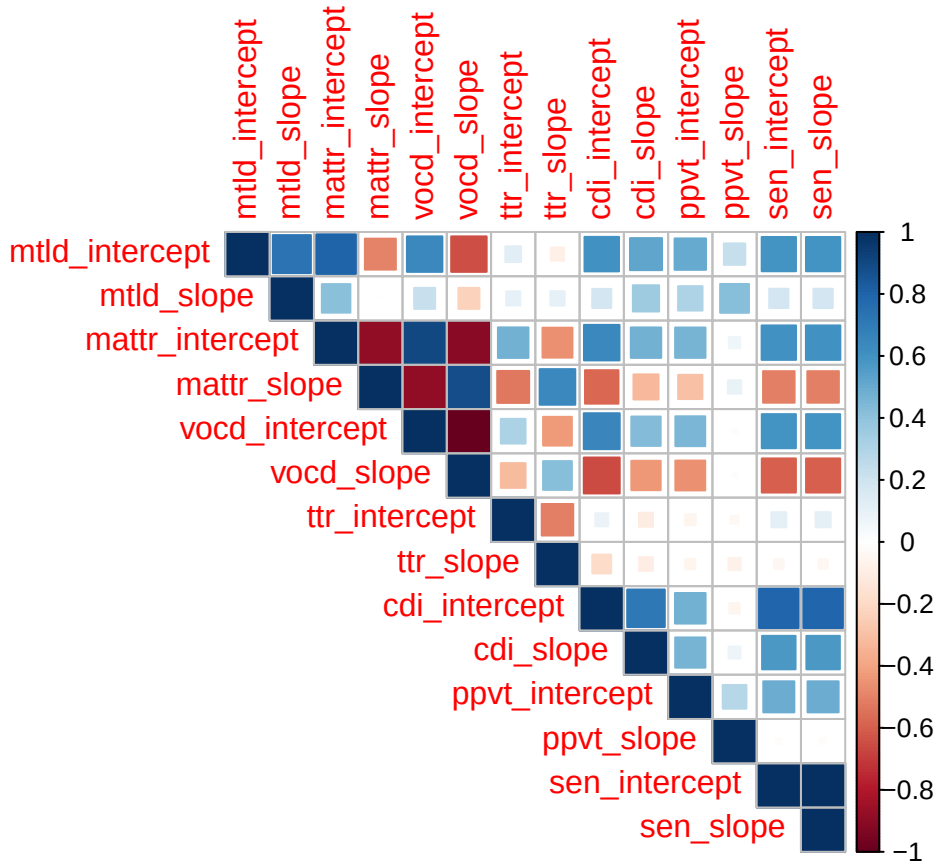


Figure 4

of caregiver's parameters as predictors of child's parameters in LDP corpus. Caregiver's intercept siginificantly relates to child's initial vocabulary skills; caregiver's slope also has a siginificant relation to child's slope (Figure 6). Notably, similar to the negative correlation of child's intercept and slope in CHILDES corpus, their caregiver's intercept also negatively relates to the slope. Correlation analysis conducted on caregivers in LDP sample reveals exactly parallel results to children.

```
## NULL
```

```
## NULL
```

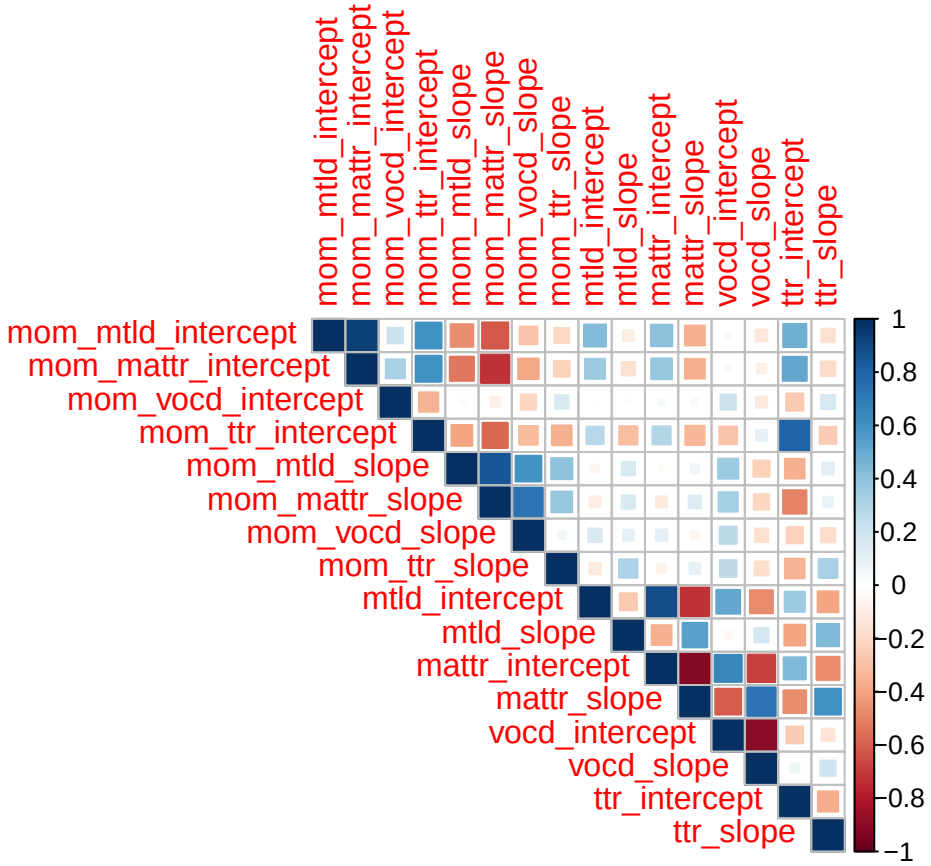


Figure 5

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 to what extent caregivers change the way they talk across age significantly relates to the growth rate of child's vocabulary, as measured by MTLD and normative measures. So far, the results generated by MTLD are consistent with the previous findings, revealing a significant relation between caregiver's speech and child's language development. 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.

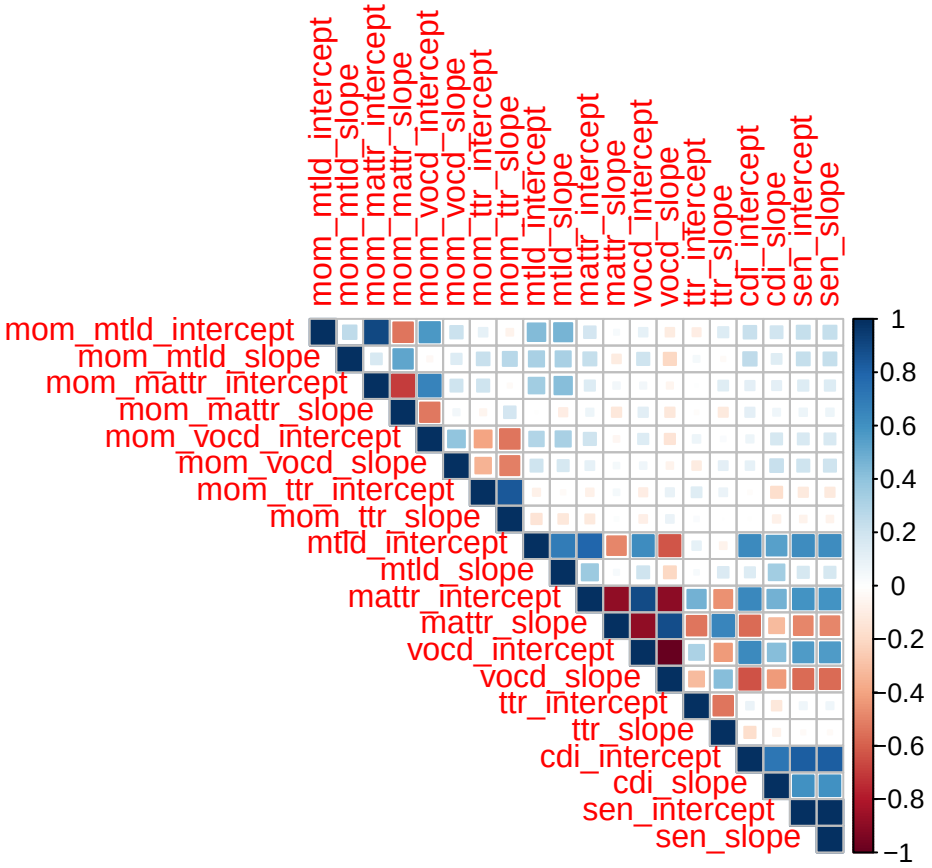


Figure 6

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 evaluated sequentially for its TTR. For example, "I"(TTR = 1) "had"(TTR = 1) "chicken"(TTR = 1) "and" (TTR = 1) "I" (TTR = 0.8) "also" (TTR = 0.83) "had" (TTR = 0.71) and so forth. When the default TTR score is reached (here, 0.72), the factor count increases by a value of 1 and the TTR evaluations are reset. This process is repeated until the last token of the sample is evaluated for its TTR. Then the total number of tokens is divided by the total factor count. Subsequently, the same process is repeated on the reversed language sample. The final MTLD value is the mean of forward and reversed MTLD scores.

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

common approach. One reason of its being ubiquitous relates to the advantage of avoiding local clustering. However, it may lead to a distorted way of overall text (Malvern et al. 2004). MTLD is an exception. The sequential analysis of MTLD distinguishes itself from other measures by maintaining the integrity of a text, because it evaluates words in order, rather than treats a text as a bag of words. Words, or other textual components, have to be bound together with a certain structure so that a reader or a listener can form a coherent mental representation (Van Dijk & Kintsch, 1983). Therefore, the sequential analysis may provide information on vocabulary from various levels, lexical level and sematic level, that interact in an intricate way. The final set of analyses explore how MTLD works differently from other measures by assessing multiple simulated text sampled from LDP.

Simulated speech

The sequential analysis differs from nonsequential analysis mainly in its measuring a text in order. We first evaluate whether the order of words in a text influences lexical diversity score of the entire text. To test this, we begin with a baseline sample of 3000 tokens from large LDP corpus and then create another two simulated child speech samples generated by including 15 tokens in a repetitive order or in a random order. For the 15 tokens, we generate a list of all the unique word types produced by children in the entire corpus, and 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 by four lexical diversity techniques. Results are shown in Figure 4. Whereas vocd-D and TTR scores decrease when adding 15 tokens into the baseline sample, regardless of orders of 15 tokens, 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.

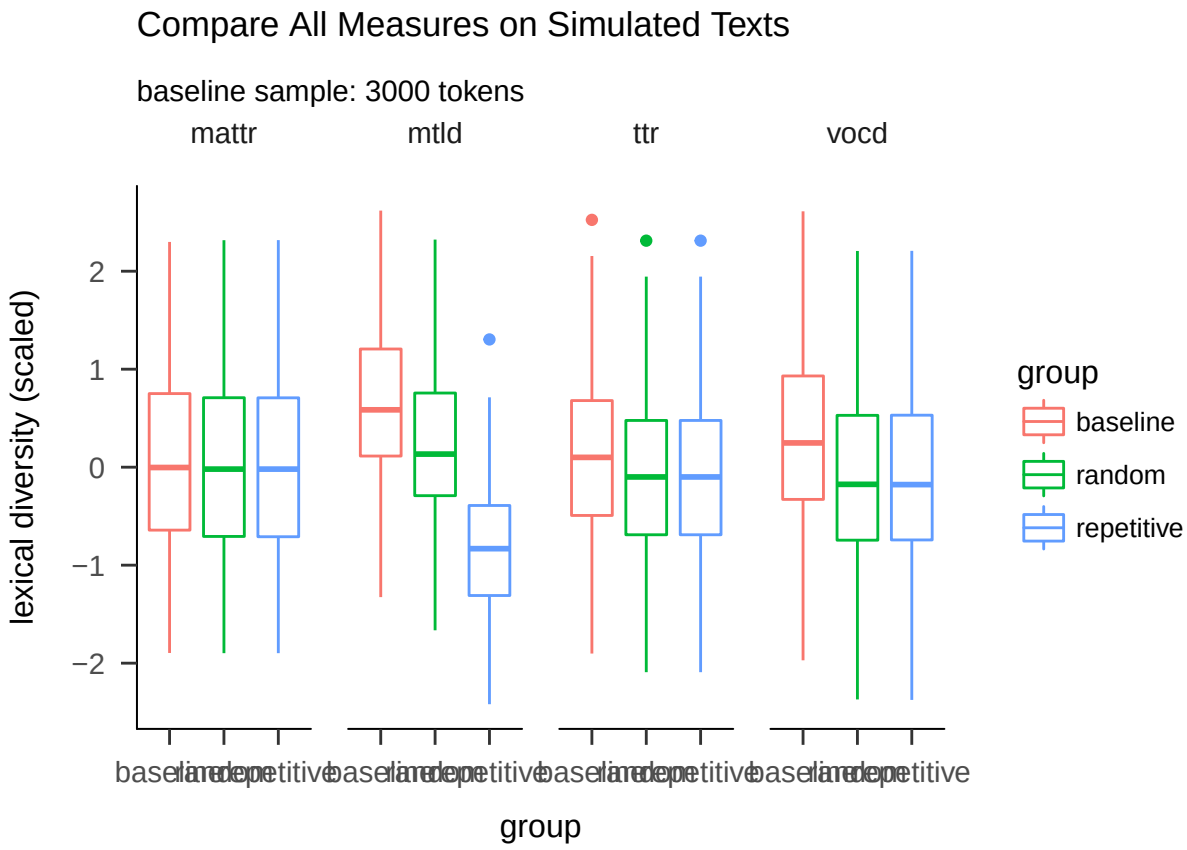


Figure 7

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 randomly sample 3000 tokens as a basline child speech and add 5 unique low-frequency word types in a repetitive order or in a random order, respectively. To be more specific, these word types are “treatment”, “clog”, “trustworthy”, “thief” and “tofu”; each word type only occurs once in the LDP corpus. The second sample comprises of the baseline sample with these 5 unique word types repeating 3 times in order, and the third sample entails these 5 word types repeating 3 times in a random order. We perform the same sampling procedure described previously 100 times. Figure 5 demonstrates MTLD scores are sensitive to word

order regardless of the frequency of words added. Whereas MATTR and TTR are influenced by neither word frequency or word order, vocd-D scores slightly increase when low-frequency words are added but decrease when adding high-frequency words.

8-1.pdf

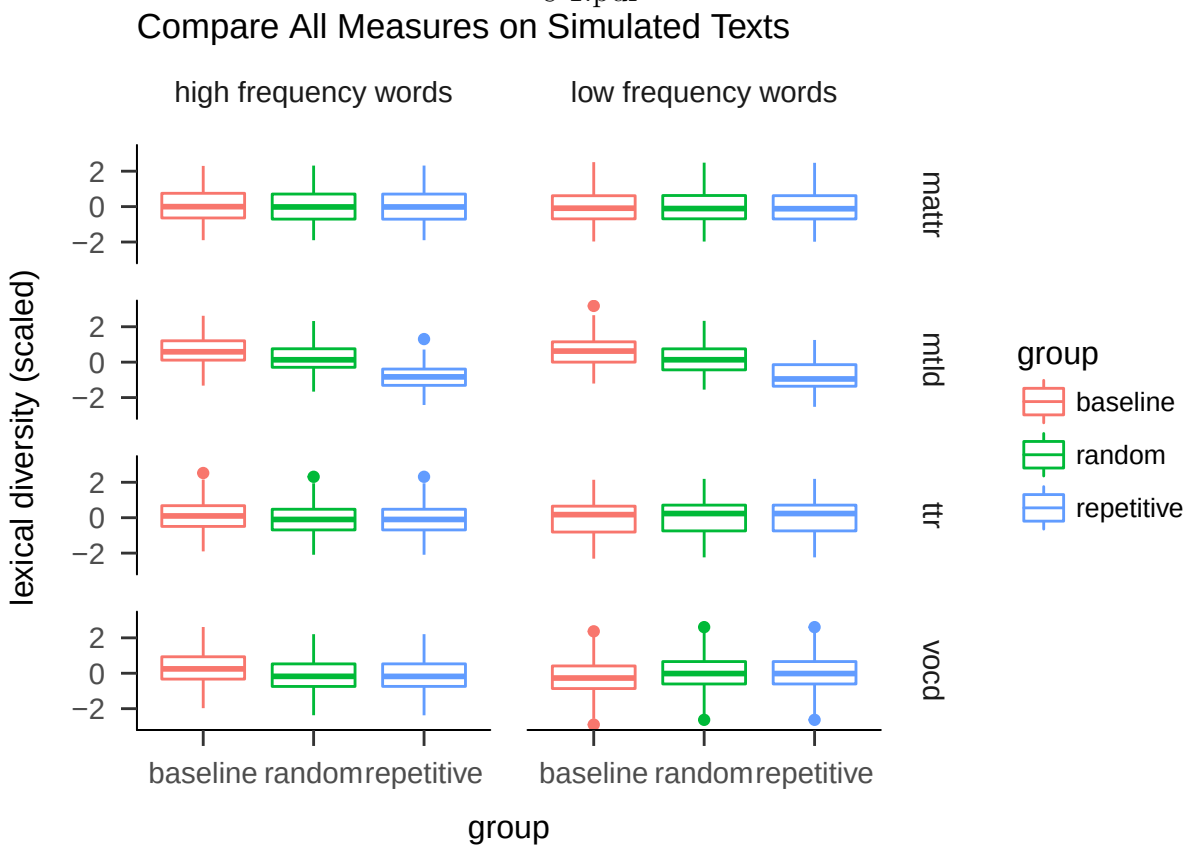


Figure 8

Discussion

Acknowledgements

We are grateful to the members of the Communication and Learning Lab for feedback on this project and manuscript. This work was supported by a James S. McDonnell Foundation Scholar Award to DY.

References