Lexical Diversity and Language 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

Large variability in quantity of linguistic input to children, but also variability in quality. In some cases this quality appears to vary across groups, in others not. What is the right meausure of quality, and how is it related to language acquisition? In addition, how does the structure of input change over development. We look at a large, diverse, longitudinal corpus to answer these questions.

Keywords: cognitive development; language acquisition; lexical diversity

Word count: X

Lexical Diversity and Language Development

```
## # A tibble: 40 x 2
##    subject lesion
##      <int> <chr>
##  1     100 <NA>
##  2     102 <NA>
##  3     103 <NA>
##  4     105 <NA>
##  5     106 <NA>
##  6     107 <NA>
##  7     108 <NA>
##  8     109 <NA>
##  9     110 <NA>
## 10     119 <NA>
## # ... with 30 more rows
```

Every typically developing child acquires language. Children learn language no matter what country they are born in or what language is spoken around them. They learn language no matter what cultural beliefs about language learning and transmission are held by the adults in their community (Lenneberg, 1967). But this universal capacity to learn belies tremendous variability in both the rates and outcomes of learning.

Some of this variability is due to differences between languages. For instance, across languages, differences in both structure and cultural practices predict different trajectories of acqusition. For instance, children learning English many other languageacross languages, children tend to acquire nouns like "ball" before verbs like "throw" (Gentner, 1982). However, this tendency appears weaker in children learning Mandarin (Tardif, 1996). One potential explanation for this difference is that Mandarin speaking caregivers talk to their children more about relations, and less about objects (Tardif, Gelman, & Xu, 1999). When these children enter school and begin leanring arithmetic, English learning children will have more trouble than Mandarin learning children in part because of the structure of the number words in their languages. The English-learning children will struggle with the teens, which are idiosyncratic and opaque relative to the words for the same numbers in Mandarin (Ho & Fuson, 1998).

But, much of the variability occurss within language across children.

Language learning is highly similar across children, contexts, languages, etc. But, language learning is also variable across children–different languages show some different orderings, some kids are slower than others, etc. How do we think about the sources of these differences? One possiblity is certainly genetic differences, but even these estimates suggest that large amounts of variability are environmental. So, how do we think about environmental differences? Lots of evidence that more is good, but not all input is created the same What is the right way of measuring quality? Lexical diversity and friends Why TTR is bad But what is the matter with TTR? - length confounds, but also context confounds Some solutions: MATTR, VOCD, MTLD differences/similarities maybe the redfish bluefish and jabberwocky example? We want to solve two problems: 1. How do we measure diversity correctly? 2. How are parents and kids related This is a chicken and egg problem. We try bootstrap our way in by looking at these different measures in their parent-child correlation and also correlation with external measures Desiderata 1. Individual parents and children are related (either by genetics or input) 2. Slope and intercept are probably related (see other rich get richer effects) 3. External validity We conclude that MTLD is the best measure, and that you get sensitivity from parents This is interesting because it suggests that we don't want a pure diversity measure, we want something in the secret sauce of MTLD. What might that be?

Methods

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

Participants

Participants in this study were parents and children drawn from the Language Development Project (LDP) – a longitudinal corpus of naturalistic parent-child interactions in a diverse sample of the Chicagoland community (Goldin-Meadow et al., 2014). We selected as our sample 66 typically developing children. Children in the corpus were recorded in their homes for 90 minutes at a time every over the age of 14 months and `ldp_all %>% pull(age) %>% max() + 1` months.

From the full set of 116 children, we excluded those who did not meet a set of criteria. First, we excluded 67 atypically developing children. We then excluded 2 children for whom we did not have at least 5 home visits in which they spoke at least 100 words per visit in order to accurate estimate individual vocabulary growth.

Linguistic data

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 66 primary caregiver-child dyads. LDP corpus contains a total of about 7 million tokens after removing a number of special transcription characters and other artifacts of the CHILDES coding system, as well as un-transcribable sections .

External measures of vocabulary size and sentence complexity

Children‘s vocabulary skills were evaluated with the use of MacArthur-Bates Communicative Development Inventories (CDIs) at 14 months and Peabody Picture Vocabulary Test (PPVT) <NOTE: DO WE KNOW WHICH VERSION> 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. In addition, MLU is computed

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 of children's language outcome and the role of caregiver's input on this outcome.

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.

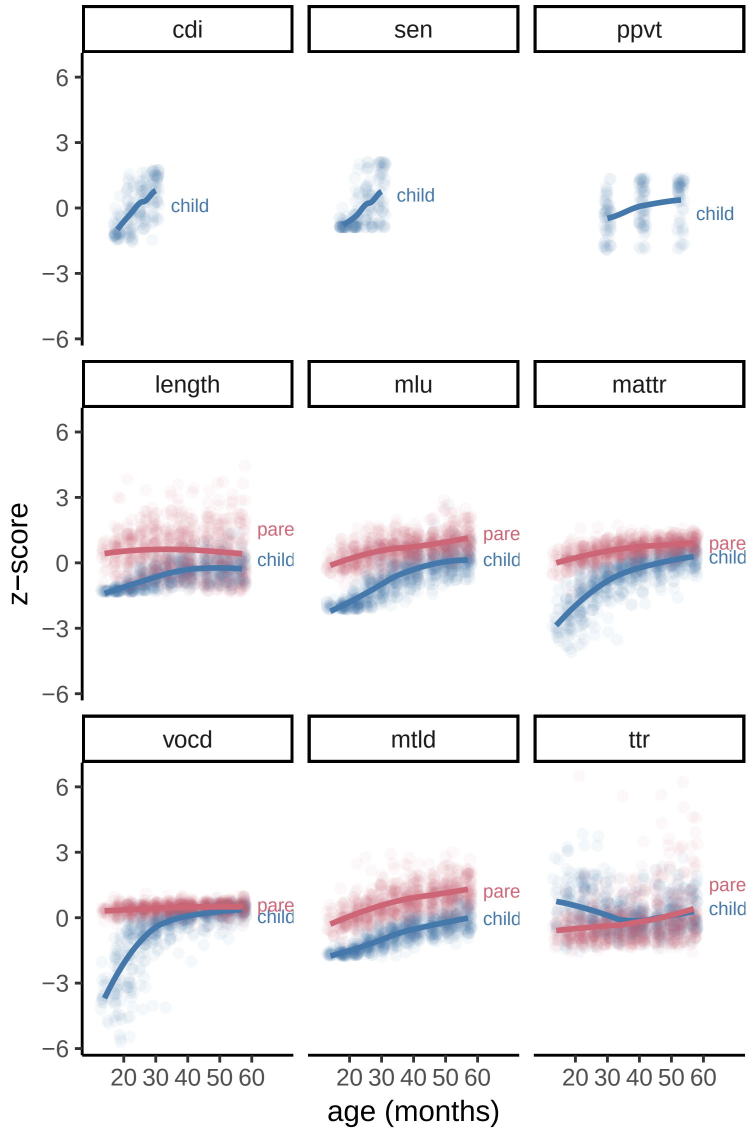
Another advantage of mixed-effects model is that local dependencies between the successive measures, specifically, vocabulary skills in preceding sessions, can be brought into the model. Lastly, it is particularly useful for handling situations in which measures for some individuals are missing at some 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 predicted based on the mixed-effects model. 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 language richness change over time.

The second question is whether there are significant individual differences among participants in LDP corpus. 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, @hoff2003, @huttenlocher1991, @huttenlocher2010, @pan2005, @rowe2008). 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 under investigation to that of normative measures, including PPVT, CDI vocabulary and CDI sentence complexity measures.

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 LDP corpus generated by MTLD, MATTR and vocd-D, that are characterized by a log-linear shape. We also plot the curves of PPVT, MLU, CDI vocabulary and sentence complexity as external norms. CDI assessments are conducted at child's early age, specifically, 18, 22, 26 and 30 months, while PPTV are conducted at 30, 41 and 53 months. They combine to represent a growth trajectory from 18 to 53 months, that lies within a specific peiord of time(i.e.14 and 58 months) intended for investigation. All measures, except for TTR curve, increase from 14 to 58 months and growth gradually diminishes over time for vocd-D.



We confirmed these visual intuitions in a set of mixed-effects models, predicting the z-scored value for each measure as a function of the log of age. A random intercept was included for each participant as random slopes did not converge (`value ~ log(age) + (1|subj)`). All of the measures of children showed significant increases with age (minimum slope = 0.96, p $< <$ .001 except for Type Token Ratio, which showed significantly decreasing slope over development (slope = -0.43, p $< <$ .001).

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 hardly represent the growth pattern of child's lexicon over time. So far, the results concur with findings in many previous research [(**???**), (**???**); Arnaud, 1984 <NOTE: what is this>, (**???**), Montag, Jones, and Smith (n.d.)) 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 MLU, with log-transformed age as a sole predictor. We obtain paramters of growth trajectories, specifically, the intercept describing initial aptitude for lexical diversity 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 varied widely in the initial vocabulary skills and the results generated by all measures are consistent, however, the variance of slope siginificantly differs with respect to various measures. For example, the largest variation in the slope is measured by type-token ratio, that is approximately 10 times as the child's slope drew from MTLD. The third goal of the study is to evaluate predictors of growth parameters of child's lexical development.

Table 1
Descriptive statistics for LDP child's intercept and growth rate (n=66)

measure	type	mean	sd	CV
cdi	intercept	0.00	349.70	241,633,433,054.82
cdi	slope	834.32	130.35	0.16
mattr	intercept	0.00	0.20	-228,731,466,525.02
mattr	slope	0.23	0.05	0.22
mlu	intercept	0.00	0.45	-536,727,528,012.13
mlu	slope	2.35	0.22	0.09
mtld	intercept	0.00	8.73	-297,795,238,104.14
mtld	slope	17.95	2.97	0.17
ppvt	intercept	0.00	53.95	1,322,653,736,213.04
ppvt	slope	50.44	16.56	0.33
sen	intercept	0.00	53.61	390,660,391,670.31
sen	slope	39.00	19.08	0.49
ttr	intercept	0.00	0.17	687,521,974,313.23
ttr	slope	-0.03	0.05	-1.66
vocd	intercept	0.00	11.38	-428,623,247,300.13
vocd	slope	10.39	2.84	0.27

Note. CV reported here is the coeficient of variation that shows the extent of variability in relation to the mean.

Correlation with maternal language and family income

Children vary widely in their intercept and slope of vocabulary growth trajectory. We first evaluate predictors of child's growth parameters generated by MTLD, MATTR, vocd-D and TTR. A growing body of previous work demonstrates significant influence of caregiver's speech on child's language development (Rowe, 2008;…). In LDP corpus, caregivers' intercept does not relate to their child's intercept, as shown in Table 2. Because the initial language aptitude is represented by one of growth parameters-intercept-we seprate confounding contextual relation from caregiver-child conversation sample. Yet, caregiver's slope significantly relates to child's growth rate of vocabulary diversity. Table 2 demonstrates a positive relation between caregiver's slope and child's slope that are deriving from MTLD and MLU, while MATTR generates a moderately negative relation between

Table 2
Correlation of Lexical Diversity Growth Rate between Mother and Child

measure	type	correlation
mattr	intercept	-0.09
mattr	slope	-0.13
mlu	intercept	-0.13
mlu	slope	0.17
mtld	intercept	0.06
mtld	slope	0.30
ttr	intercept	0.01
ttr	slope	0.03
vocd	intercept	0.03
vocd	slope	-0.07

them. This finding aligns with that from previous work in which mothers fine-tune language usage in connect to their children's level of understanding and language skills.(. . .)

Score of research documents a relation between socioeconomic status and children's vocabulary development (Hart & Risley, 1995; Lawrence & Shipley, 1996; Hoff-Ginsberg, 1991; Hoff, Laursen & Tardif, 2002). Our results also show household income is a significant predictor of children's lexicon diversity. Table 3 presents a significant correlation between household income and child's intercept generated by MTLD, CDI, PPVT and MLU, and its relation to child's slope measured by MTLD, MLU and PPVT. Specifically, children of high SES do not necessarily start with a more sophiscated language skill, but their vocabulary tend to develop faster than children from lower income family.

The mechanism of MTLD

The correlation analysis demonstrates that household income is a significant predictor of child's vocabulary skill, and to what extent caregivers change the way they talk across age

Table 3
Correlation between Child's Lexical Diversity and Family Income

measure	intercept	slope
cdi	-0.11	0.07
mattr	-0.03	0.05
mlu	-0.24	0.24
mtld	-0.28	0.30
ppvt	-0.23	0.39
sen	-0.02	0.02
ttr	-0.03	-0.01
vocd	0.01	-0.01

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.

Sequential analysis

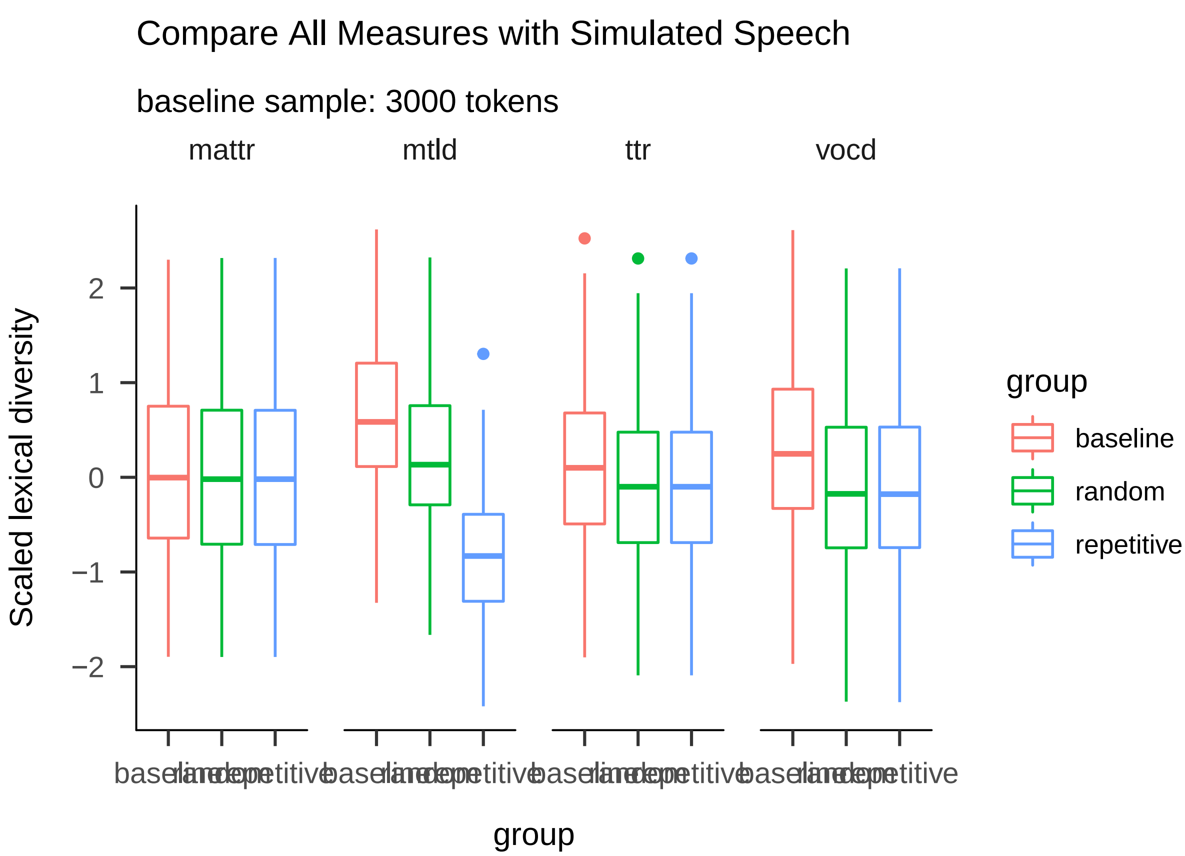
Conceptually, MTLD estimates average number of consecutive tokens for which a certain TTR is maintained (e.g. 0.72 by default). For any given sample, each token is evaluated sequentially for its TTR. For example, "I"(TTR = 1) "had"(TTR = 1) "chicken"(TTR = 1) "and" (TTR = 1) "I" (TTR = 0.8) "also" (TTR = 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 child's speech sampled from LDP corpus.

Simulated speech

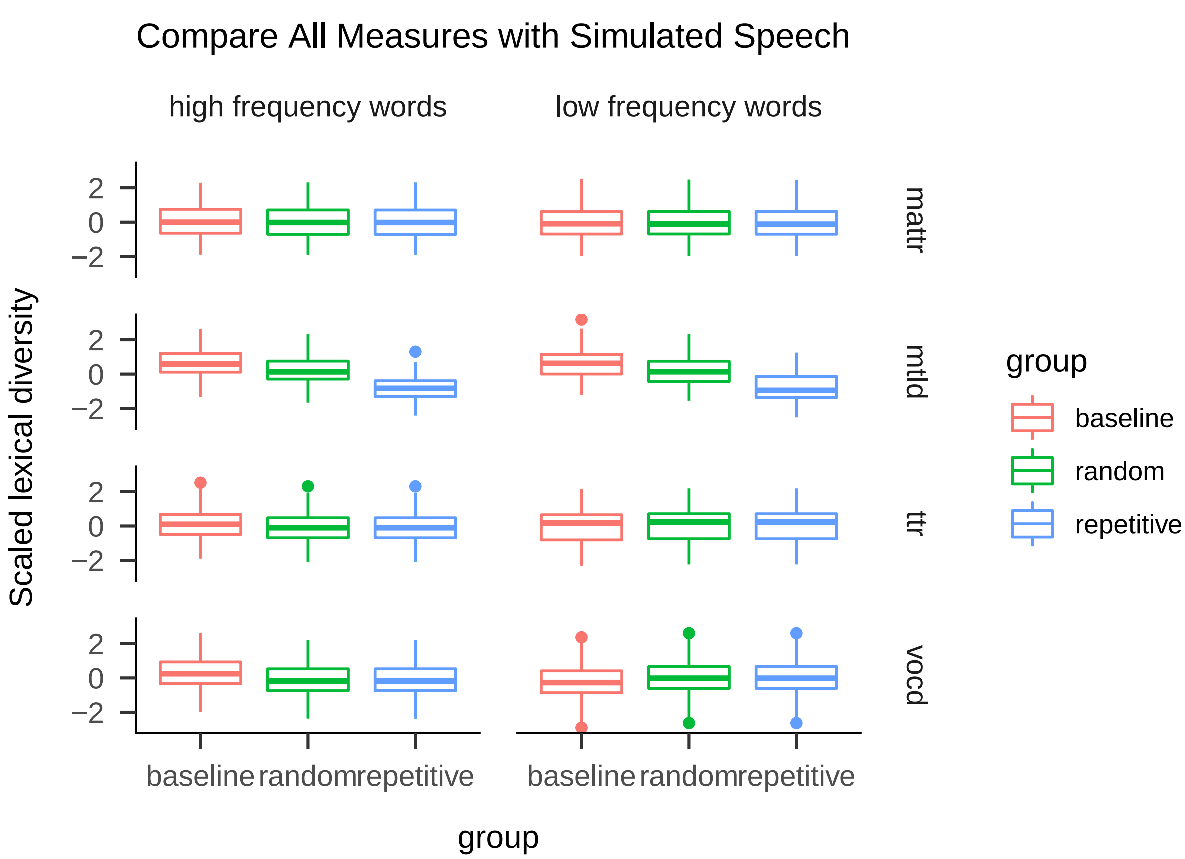
The sequential analysis differs from nonsequential analysis mainly in its measuring a text in order. Here, we sougt to assess the degree to which there is a significant change in the value of each lexical diversity index caused by the change of word order. We begin with a baseline sample of 3000 tokens from 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 versions of child speech by four lexical diversity techniques.

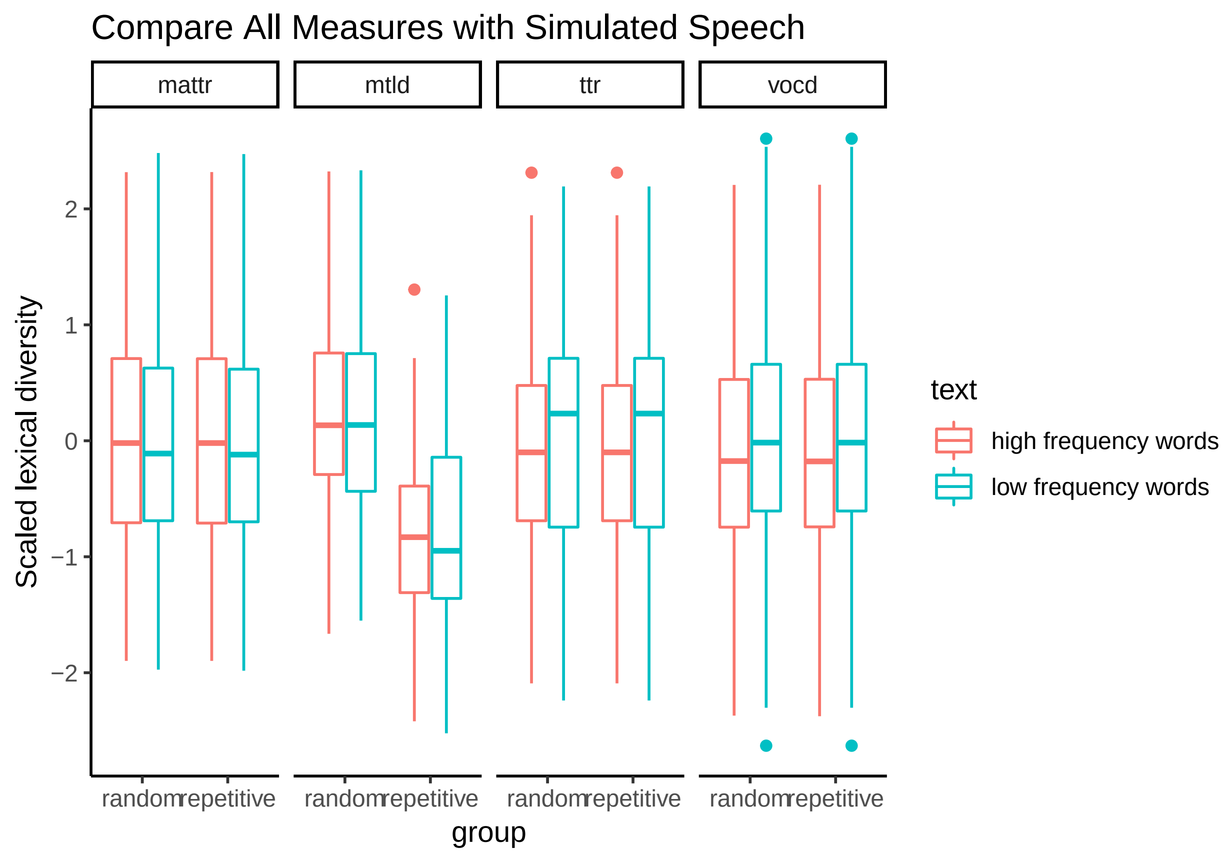
Results are shown in Figure 2. There is a consistent decrease in MTLD scores when comparing samples of different word orders, though only 0.5 percent of tokens are manipulated. Whereas MATTR shows no change in its value with any manipulation, vocd-D and TTR scores slightly decrease as 15 tokens are added into baseline sample, regardless of the word order. However, it remains unclear whether the decrease in MTLD scores is caused by the change of word orders, or adding frequent word types that actually yields greater lexical overlap. Similarly, it is also unknown if the change of vocd-D and TTR values are caused by less diversity in word types or confounded by change of text length.



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 and 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 entire 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 3 demonstrates that MTLD scores significantly drop when adding tokens in a repetitive order, but there is no significant change with various word frequencies. Whereas MATTR and TTR are influenced neither by word type nor by word order, vocd-D scores slightly increase as sparsely occuring words are added but decrease when adding more common words. Comparisons among four versions of manipulated speech of the same text length (i.e. 3015 tokens) suggest that the sensitivity of vocd-D to word types and the sentivity of MTLD to word orders are not confounded by text length.





Discussion

Previous research has shown large individual difference in children's language skills and the rate of their language growth (Huttenlocher, Haight, Bryk, Selzer, & Lyons, 1991; Fenson et.al.,1994; Hart & Risley, 1995) relating to quantity and quality of language input (Hoff, 2003; Huttenlocher et al., 1991), that vary in SES (Hess & Shipman, 1965; Heath, 1983; Hoff-Ginsberg, 1990; Hart & Risley, 1995; Rowe, 2008). This study also documented the large variation in children's vocabulary diversity and its relation with maternal language and SES (i.e. family income). However, until our study, little evidence has been presented regarding how this variation and relation differ with respect to various measures. Different language measures have generated different results relating to the variation in individual vocabulary diversity and the rate of their lexcial diversity growth as well as their relation to maternal language.

The findings from our study has made it clear that the heterogeneity of child's language skill is contingent on how it is measured, in addition to the environmental factors discussed above. MTLD is the only lexical diversity indice that has detected the positive correlation between children's language outcome and parental language input, that are consistent with the results yielded by external normative measures (i.e. PPVT and CDI) and exisiting literature.

The assessment of simulated speech/text provides evidence that MTLD and vocd-D captures different information of lexical diversity. MacCarthy & Jarvis (2010) demonstrates that lexical diversity indices cannot be assumed to evaluate the same latent trait. This study takes a step futher determining the specificity of information each measure captures. For example, vocd-D shows high sensitivity to word types, thus offering an incremental advantage of assessing child's vocabulary size, whereas MTLD distinguishes word orders, thus offering synatic and grammatical information of child's language usage.

The unique information apprehended by the two measures together delineate the construct of lexical diversity skills, though far from comprehensive. This study explores the specific aspect of child's language development between 14 and 58 months. The time range examined here grasps the linguistic transitions from producing first word to succesive single-word utterances then to meaningful sentences. Such a transition requires more than just expanding vocabulary size, but comprehending the relations between single words and which words can be meaningfully combined in what order (MacWhinney, 2011). Language development, as described by lexical diversity trajectory, is not a linear process. Rather, it can be likened to a tapestry composed of many different colors stands (i.e. phonological, lexical, semantic and syntactic skills etc.) and it can only be properly viewed and understood as a totality. Partial examination of any given section, such as vocabulary size assessment, of the tapestry yields merely an accumulation of its compotent (Irwine &

Mitchell, 1986). Future work is needed to better determine if the specific information captured by these measures varies with different registers. As such, moving the holistic understanding of child's language acquisition forward requires researchers to fully appreciate the mechanism undepining each language measure and be aware of limitations and advantages of each approach.

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

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 125–127.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; No. 257.*

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*, 588–599.

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

Ho, C. S.-H., & Fuson, K. C. (1998). Children's knowledge of teen quantities as tens and ones: Comparisons of chinese, british, and american kindergartners. *Journal of Educational Psychology*, *90*(3), 536.

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5), 1368–1378.

Huttenlocher, J., Haight, W., Byrk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*, 236–248.

Huttenlocher, W., J. (2010). Sources of variability in children's language growth. *Cognitive Psychology.*

Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, *2*(12), 59–67.

Montag, J. L., Jones, M. N., & Smith, L. B. (n.d.). Quantity and diversity: Simulating early word learning environments. *Cognitive Science*.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal Correlates of Growth in Toddler Vocabulary Production in Low-Income Families. *Child Development*, *76*(4), 763–782.

Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*, 439–421.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. *Developmental Psychology*, *32*(3), 492.

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of english and mandarin. *Child Development*, *70*(3), 620–635.