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 146 typically developing children over the age of 14 months and 58 months.

The first database we used 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 61 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.

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 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 146 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 caregiver's input influence 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 of 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 the log 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 four 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 issue 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 of both child and caregiver. 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 individual differences, the third question to be answered is whether these differences vary with different lexical diversity measures (e.g. MTLD, TTR, vocd-D and MATTR). The parameters deriving from these lexical diversity indices are also compared to that of normative measures, including PPVT, CDI and the mean length of utterance(MLU). Mean length of utterances in words (MLUw) is calculated in two corpora. Though mean length of utterances in morphemes(MLUm) are more widely used and accepted, research has found MLUw is perfectly correlated with MLUm and can be equally effectively used as a measurement of child's gross language development (Parker& Brorson, 2005).

The fourth question is whether any growth curve parameters of children can be predicted by the diversity of speech and its change of their caregivers over time. 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. Here, we utilized mixed-effects model to examine correlations between intercepts and slopes of children and their caregivers with respect to various measures.

Results

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