Analyzing Vocabulary Commonality Index Using Large-scaled Database of Child Language Development

Yan Cao[1,2], Yasuhiro Minami[1], Yuko Okumura[2], and Tessei Kobayashi[2]
[1]The University of Electro-Communications, [2]NTT Communication Science Laboratories
[1]1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan, [2]2-4 Hikaridai, Seika, Kyoto 619-0237, Japan
caoyan@sd.is.uec.ac.jp, minami.yasuhiro@is.uec.ac.jp, {okumura.yuko, kobayashi.tessei}@lab.ntt.co.jp

Abstract

The present study proposed a vocabulary commonality index for child language development to investigate to what extent each child acquires common words during the early stages of lexical development. We used large-scaled, vocabulary-checklist data from Japanese-speaking children (N=1,451) aged 8-48 months to estimate their age of acquisition (AoA) of 2688 words by logistic regression. Then we calculated the vocabulary commonality index for each child with two datasets. The results showed that as their vocabulary size increases, children who have the same vocabulary size tend to produce common words with the same ratio.

Keywords: Large-scaled Vocabulary Database, Child Language Development, Age of Acquisition, Vocabulary Commonality Index

1. Introduction

How children acquire vocabulary is among the most central issues in the fields of cognitive science and developmental psychology. Many previous studies have scrutinized what types of words young children acquire during their early stages of lexical development using a vocabulary checklist methodology, such as the MacArthur-Bates Communicative Development Inventories (e.g., Bates et al., 1995; Caselli et al., 1995; Fenson et al., 1994; Frank et al., 2017). For example, Caselli, Casadio and Bates (1999) found that 18-30 month-old English- and Italian-speaking children tend to produce more social words (e.g., people's names, games, routines, etc.) than other types of words in their first 50 words and more common nouns (e.g., animals, foods, toys, etc.) in their first 100-500 words. Although such category analyses of vocabulary are useful for grasping children's overall tendencies in developmental stages, they are less satisfactory for understanding more detailed changes and the individual differences of vocabulary development because the category ranges are wide-ranging and sparse (i.e., only four categories).

One possible solution for grasping detailed changes and individual differences in vocabulary development is to directly compare the word items (rather than categories) acquired by each child with the word items acquired on average by many children and roughly estimate the commonality between both word sets. If we can successfully estimate the commonality in a statistically reliable way, we might be able to understand to what extent each child acquires common words and further clarify the detailed changes and individual differences of vocabulary development.

In the present study, we propose a vocabulary commonality index for child language development by estimating to what extent each child can say common words. We also create a new vocabulary-checklist that includes much more word items (i.e., 2688 words) than the Japanese version of MacArthur-Bates CDI (i.e., 448 words). This is because we thought that the word range of MacArthur-Bates CDI was insufficient to grasp detailed changes and individual differences in vocabulary development. Therefore, using such a new vocabulary-checklist, we create a large-scaled database of vocabulary database in Japanese-speaking children.

2. Data Collection

The data for the study came from two datasets.

2.1 Dataset 1: Tablet Survey

2.1.1 Participants

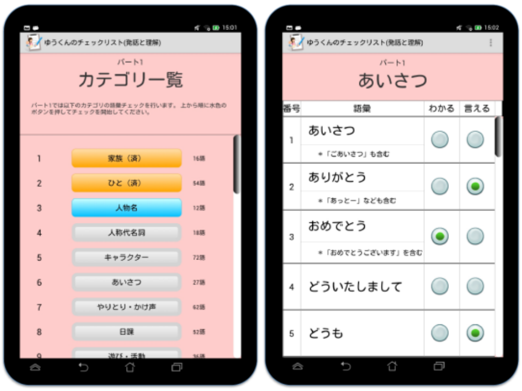
We collected the data of 1,451 Japanese-speaking toddlers/children (776 boys and 675 girls) whose ages ranged from 8 to 48 months from their parents living in Kyoto, Osaka, and Nara. The participants were recruited from a local newspaper.

2.1.2 Vocabulary Checklist Application

We collected vocabulary data from parents using a tablet PC application that included 2,688 words (Kobayashi, Okumura & Minami, 2016). First, we selected 2,052 words as basic words (common nouns, verbs, adjectives, etc.) from the early vocabulary list that we longitudinally collected from about 800 web users throughout Japan (Kobayashi & Nagata, 2010). Next we added 636 such special words as anime characters and railroad/train names. These 2,688 word items are the most likely ones acquired by children who grew up in a Japanese environment and included almost all of the words from the Japanese version of MacArthur-Bates CDI (Ogura, & Watamaki, 2004; Watamaki & Ogura, 2004).

At the child playroom in our laboratory, the participants checked whether their child could comprehend or say each word (Fig. 1). The vocabulary checklist application consisted of three parts. In Part 1, 2,052 basic words were classified into 29 categories (Fig. 2). The participants had to complete all of the categories before they could finish Part 1 and move to Part 2, which consisted of 636 special words. Part 3 consisted of a free description field that allowed participants to add words that were not on the checklist.

Fig. 1 Parent using tablet PC in a playroom



ゆうくんのチェックリスト(発話と理解)

パート1

カテゴリ一覧

パート1では以下のカテゴリの語彙チェックを行います。上から順に水色のボタンを押してチェックを開始してください。

1	家族（済）	16語
2	ひと（済）	54語
3	人物名	12語
4	人称代名詞	18語
5	キャラクター	72語
6	あいさつ	27語
7	やりとり・かけ声	62語
8	日課	52語
9	遊び・活動	36語

ゆうくんのチェックリスト(発話と理解)

パート1

あいさつ

番号	語彙	わかる	言える
1	あいさつ ＊「ごあいさつ」も含む	○	○
2	ありがとう ＊「あっとー」なども含む	○	●
3	おめでとう ＊「おめでとうございます」を含む	●	○
4	どういたしまして	○	○
5	どうも	○	●

Fig. 2 Vocabulary checklist application: category list in Part 1 (left) and a panel to check comprehension or production to target words (right)

2.2 Dataset 2: Web Survey

2.2.1 Participants

We sent out an in-depth survey on children's vocabulary acquisition to the members of a large internet research company called Macromill. We received responses from the parents of 1,446 Japanese-speaking toddlers/children (684 boys, 759 girls and 3 with no reported sex) whose ages ranged from 8 to 48 months.

2.2.2 Web-based Survey

Parents completed a web-based survey that asked them to check whether their child could just comprehend a word (without saying it), or whether they could say it or whether neither option fit for identical word items (N = 2,688) in Dataset 1. This survey also asked respondents to input such information as gender, age range, prefecture, region, education level, and job.

3. Methods

We used data from the Tablet Survey to estimate the age of acquisition (AoA) due to its reliability and analyzed the vocabulary commonality index with two datasets.

3.1 Age of Acquisition

As a first step, we calculated the age of acquisition (AoA) in the following way. First, we calculated the acquisition rate of comprehending and speaking each word at every month. According to Minami and Kobayashi (2013), acquisition rate *f(x)* of comprehending and speaking at every month *x* is modeled by the logistic function of Eq. (1). We also introduced parameter a that set an upper limit, which was different from the standard logistic function:

$$f(x) = \frac{ae^{cx+b}}{1 + e^{cx+b}}. \qquad (1)$$

The acquisition curves of all the words were modeled by a logistic function using a nonlinear least squares method based on the Gauss-Newton algorithm. We also set constraints so that the upper limit of the acquisition curve of comprehension exceeded the upper limit of the acquisition curve of speaking because we believe that a children's tendency to comprehend precedes speaking. Fig. 3 shows an example of the acquisition curve. The squares indicate the comprehension data, and the circles indicate the speaking data. The logistic function clearly fits in both cases.

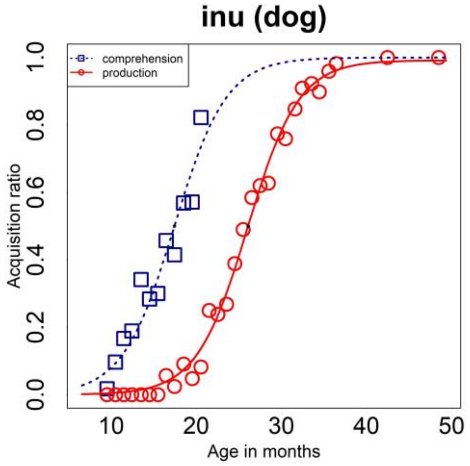


Fig. 3 Acquisition curve of inu (dog)

We also estimated the AoA for each word: the age at which 50% of children can say a word (calculated by x satisfying *f(x) = 0.5* in Eq. (1) with Brent's method). Then we listed those words by their AoA estimation orders. Table 1 shows the first 30 words from the AoA estimation.

3.2 Vocabulary Commonality Index

To clarify the developmental pattern of child vocabulary growth, we estimated to what extent each child can say common words, i.e., how child vocabulary commonality changes based on developmental levels. The Vocabulary Commonality Index (*VocIndex*) is calculated as follows in Eq. (2):

$$VocIndex(i) = \frac{|\, set\ Voc(i) \cap set\ AoA(i)\, |}{|\, set\ AoA(i)\, |}, \qquad (2)$$

where $set\ Voc(i)$ is a set of words that child *i* can say and $set\ AoA(i)$ is a set of words ranked by AoA estimation order with the same number of words in $set\ Voc(i)$. $|\, set\ AoA(i)\, |$ counts how many words are in $AoA(i)$. Moreover, from Eq. (2), note that the higher the vocabulary commonality index is, the more common words a child can say. If all the words that a child can say are common words, her vocabulary commonality index will be 1.

4. Results and Discussion

4.1 Analysis 1: Data from Dataset 1

We calculated the vocabulary commonality index for each child and plotted two types of graphs for it. Fig. 4 shows the relationship between the vocabulary commonality index and age in months. Each point represents an individual child, indicating their age in months and vocabulary commonality index. Because the distribution of the scatter plots was irregular, no clear relationship seems to exist between the vocabulary commonality index and a child's age in months.

Table 1 First 30 words in production for Japanese children based on the AoA estimation. The list includes many infant-directed speech (IDS) words.

Rank	Word	Translation	AoA (days)
1	inaiinaiba	peek-a-boo	433.9
2	manma	(IDS word of meal)	454.8
3	wanwan	(dog sound)	474.4
4	mama	mommy	481.0
5	hai	yes	506.2
6	aq	oh (expression of surprise)	509.2
7	papa	daddy	511.4
8	baibai	bye-bye	513.6
9	a-a	aah (expression of failure)	530.2
10	anpanman	(character name)	546.3
11	nenne	(IDS word of sleep)	561.1
12	bubbu	(vehicle sound)	561.3
13	nyannyan	(cat sound)	561.8
14	iya	no	565.3
15	a-n	(request for opening mouth)	597.6
16	wanwan	(character name)	601.1
17	nainai	(IDS word of cleaning)	606.5
18	ba-ba	grandma	617.6
19	kukku	(IDS word of shoes)	618.0
20	douzo	Here you are	621.3
21	pan	bread	623.2
22	dakko	(IDS word of holding)	625.0
23	shi	(pee sound)	627.5
24	arigatou	thank you	629.7
25	chu	(kiss sound)	630.7
26	un	yes	641.2
27	ji-ji	grandpa	641.2
28	ocha	tea	645.6
29	atchi	there	647.8
30	fu-fu	(blow sound)	654.9

In contrast, Fig. 5 shows a strong relationship between vocabulary commonality index and child vocabulary size. When the vocabulary size was set to the x-axis, a vocabulary commonality index pattern is greatly suggested. When the vocabulary size of each child is small, the vocabulary commonality index varies widely. However, as the vocabulary size increases, the deviation of the vocabulary commonality index gradually becomes smaller. In other words, children who have the same vocabulary size tend to produce common words with the same ratio. One possibility why the vocabulary commonality index decreases from around 2000 words is that as children's vocabulary size increases, they tend to produce more words that are not included on the vocabulary checklist.

Third, using the vocabulary commonality index makes it possible to identify any parents who randomly checked the answers in the vocabulary checklist application. We calculated the probability that common words existed in the AoA word list when the participants randomly selected the answers in the vocabulary checklist application and added the result to the red line in Fig. 5. No participants improperly completed their questionnaires.

Finally, to show the data's trend more clearly, we calculated the moving average values for the vocabulary commonality index and the vocabulary size with intervals of 60 data (Fig. 6). We also calculated the moving standard deviations with a window size of 60 on the vocabulary commonality index. As the total vocabulary size continues to increase, the vocabulary size index also increases. However, during the beginning of children's vocabulary production, the vocabulary commonality index once falls and then goes up again. Table 2 shows the moving average (MA) values and the moving standard deviations (MSDs) of some vocabulary sizes.

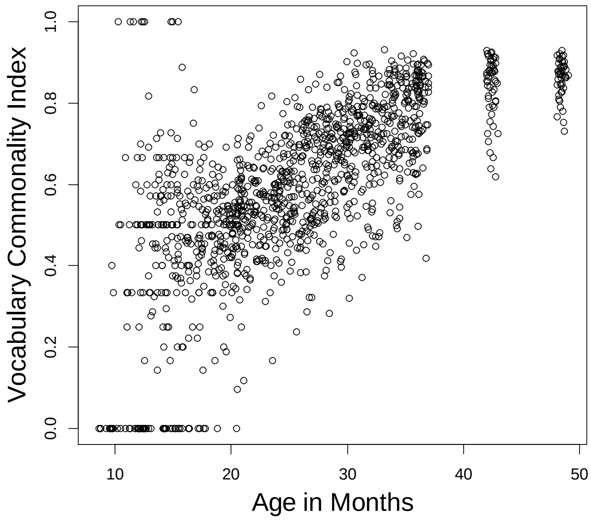


Fig. 4 Vocabulary commonality index plotted by age in months (dataset 1)

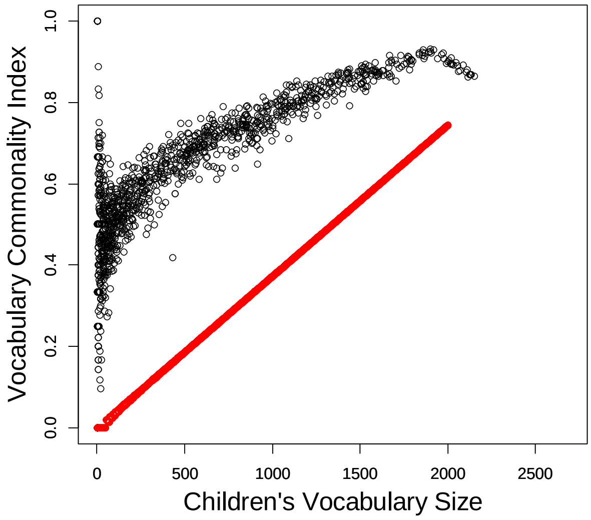


Fig. 5 Vocabulary commonality index plotted by children's vocabulary size (dataset 1)

4.2 Analysis 2: Data from Dataset 2

Using data from Dataset 2 we calculated the vocabulary commonality index for each child and plotted a scatter plot (Fig. 7). Each point shows one child's vocabulary size versus her vocabulary commonality index. Fig. 7 shows a strong relationship between the vocabulary commonality index and vocabulary size (Fig. 5). Across the Tablet and Web Surveys, as the size of their vocabulary increases, children with the same vocabulary size tend to say common words at the same ratio.

However, several outliers are markedly distant from other points in Fig. 7. To identify these outliers, we consider points under or over 2 times the standard deviation from the moving average outliers. The results are shown in Fig. 8.

We proposed a vocabulary commonality index for child language development by a mathematical method. The present results across two datasets identified the following child vocabulary development pattern: as vocabulary size increases, children at earlier stages of lexical development tend to produce common words at a certain, stable proportion. These findings may play important roles in further studies of child language development. However, the results of this study are limited to 2,688 words. Future studies need to look at the properties of the vocabulary commonality index with a smaller vocabulary size and test for a significant difference between genders.

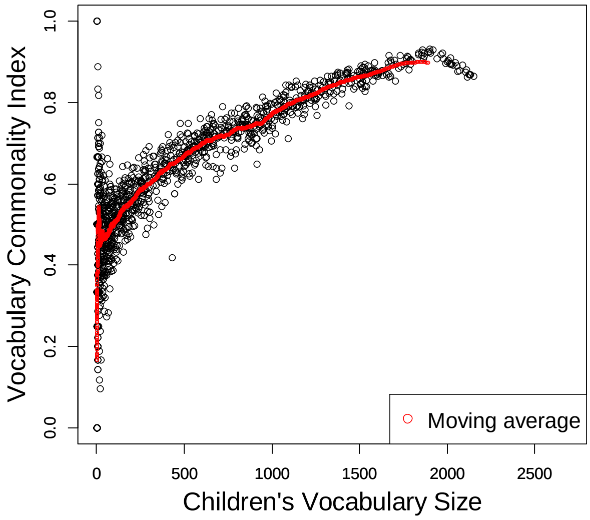


Fig. 6 Moving averages (dataset 1)

Table 2 Moving average values

Vocabulary Size	MA of VocIndex	MSD	MA + 2MSD	MA – 2MSD
30	0.47	0.10	0.67	0.27
50	0.47	0.08	0.63	0.31
70	0.48	0.08	0.64	0.32
90	0.50	0.05	0.60	0.40
150	0.54	0.05	0.64	0.44
300	0.60	0.05	0.70	0.50
450	0.65	0.05	0.75	0.55
850	0.74	0.03	0.80	0.68
1200	0.81	0.02	0.85	0.77
1850	0.90	0.02	0.94	0.86

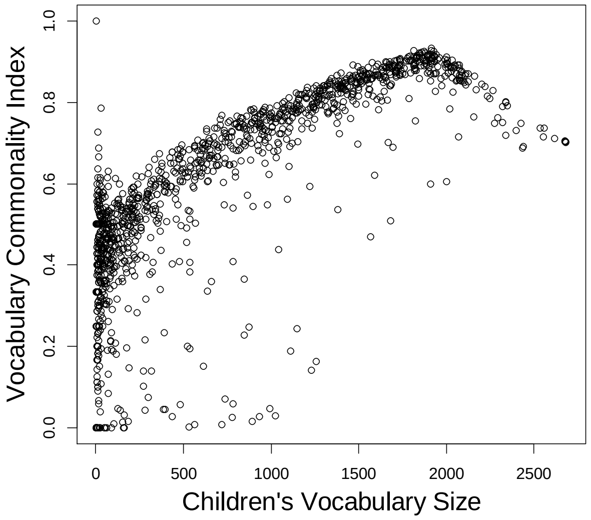


Fig. 7 Vocabulary commonality index plotted by children's vocabulary size (dataset 2)

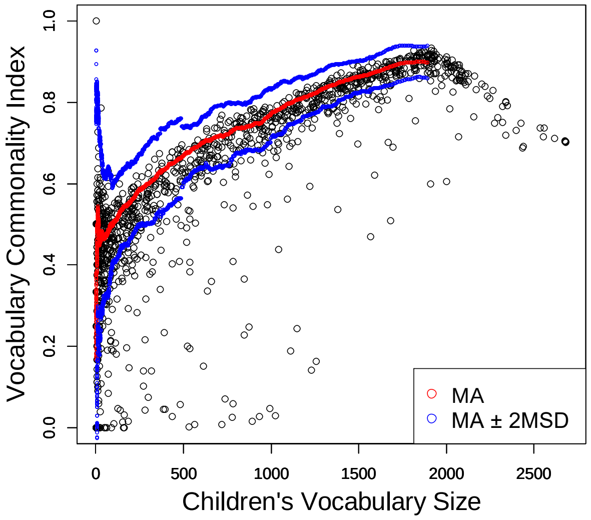


Fig. 8 Moving average (MA) and 2 times moving standard deviation (MSD) from the moving average (dataset 2)

5. Acknowledgements

A part of this study was supported by a Grant-in-Aid for Scientific Research (B) 17H02190. The authors thank the parents for participating in this study.

6. References

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. The Handbook of Child Language, 96-151.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. Cognitive Development, 10(2), 159-199.

Caselli, C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. Journal of Child Language, 26(1), 69-111.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., & Stiles, J. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, i-185.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. Journal of Child Language, 44(3), 677-694.

Kobayashi, T. & Nagata, M. (2010). Early lexical development in Japanese Children: Age of acquisition of first 50 words specified using web diary method. The 20th Child Language Seminar (CLS2010), London, UK.

Kobayashi, T., Okumura, Y., and Minami, Y. (2016). Collecting data on child vocabulary development database by vocabulary-checklist application. IEICE Technical Report HCS2015-59(2016-01), pp. 1-6.

Minami, Y. and Kobayashi, T. (2013). Developmentally-Appropriate Vocabulary Search System. The IEICE transactions information and systems D, Vol., J96-D, No.10, pp. 2612-2624.

Ogura, T., & Watamaki, T. (2004). The Japanese MacArthur communicative development inventory: Words and gestures. Kyoto: Kyoto International Social Welfare Exchange Center.

Watamaki, T., & Ogura, T. (2004). The Japanese McArthur communication development inventory: Words and sentences. Kyoto: Kyoto International Social Welfare Exchange Center.