Infant Word Comprehension-to-Production Index Applied to Investigation of Noun Learning Predominance Using Cross-lingual CDI database

Yasuhiro Minami[1], Tessei Kobayashi[2] and Yuko Okumura[2]
The University of Electro-Communications[1], NTT[2]
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan[1], 2-4 Hikaridai Seika-cho Soraku-gun, Kyoto, 619-0237, Japan[2]
minami.yasuhiro@is.uec.ac.jp[1], (kobayashi.tessei, okumura.yuko)@lab.ntt.co.jp[2]

Abstract

This paper defines a measure called the comprehension-to-production (C2P) index to investigate whether nouns have predominance over verbs in children's word learning that identifies a partial word learning period from comprehension to production. We applied the C2P index to noun predominance using cross-lingual child communicative development inventory databases and confirmed that it indicates noun predominance in word learning, suggesting that the process between a word's comprehension and its production is a significant factor of predominance in noun learning by children.

Keywords: vocabulary learning, word acquisition, cross-lingual CDI database

1. Introduction

At around one year of age, infants produce their first words and rapidly start acquiring more of them. Almost every child follows this tendency. However, the content of a child's first few words depends on the individual. These variances might reflect different culture, language, and individual environments. In the children word learning process, identifying what parts of speech children learn first is important; this result would be a step toward resolving the children's word acquiring mechanism and would provide clues for engineering solutions for teaching words to computers. Gentner argued that nouns should predominate verbs since they appear more often than verbs in the early development stages of children (Gentner, 1979; Gentner et al., 2001).

There are two hypotheses for this predominance: the existence of a word continuum and constraints (biases). Gentner explained the former hypothesis using a division of dominance continuum. She assumed a word continuum in abstract space that varies from cognitive to linguistic dominance and introduced two assumptions, natural partitions and relational relativity, to explain that children learn words in the cognitive dominance region earlier than words in the linguistic dominance region. Since the natural partitions insist that concrete objects and entities are easier to individuate in the world, children easily acquire nouns. Relational relativity insists that a verb's meaning is not isolated and that it depends on the surrounding words.

Maguire et al. explained similar reasons for noun predominance using the shape, individuation, concreteness, and imageability (SICI) continuum (Maguire et al., 2006). These factors are assigned to the one dimensional abstract space axis (from left to right). The instances of words are arranged on the axis where the far left instance has an easy shape, simple individuation, high concreteness, and high imageability. Considering the meaning of nouns and verbs, we can allocate nouns to the left hand side and verbs to the right hand side. The SICI continuum has distributions of nouns and verbs that describe their individual difficulty differences.

However, no direct evidence has connected the concrete measure of word difficulty to abstract spaces because no index, other than the occurrence rate of the part of speech, has expressed word difficulty. Since the occurrence rate is strongly affected by such environments as culture and parent input, measuring word difficulty is not appropriate. This assumption, that noun acquisition predominates verb acquisition, remains controversial (Benedict, 1979; Tardif, 1996).

This paper defines a concrete measure to evaluate noun predominance in word learning. We apply this measure to evaluate noun predominance using a cross-lingual child word development database. We also discuss the reason underlying noun predominance.

2. Crosslingual CDI databases

The MacA rthur Communicative Development Inventories (CDI) (Dale et al., 1996), which are based on parent reports, are used to check when children comprehend and produce a particular word. Mothers of children of a certain age or less complete the Words and Gestures (WG) CDI form, and mothers of children over a particular age by months fill out the Words and Sentences (WS) form. WG has two columns that verify whether the toddler understands or understands/says a particular word. Checking the "understand" column means that the child completely comprehends the word. Checking the "understand/say" column means that the child has produced the word. WS has only one column to verify whether a child can say a particular word. Checking this column means that the child has produced that word.

Even though CDI was originally develped for English (Dale et al., 1996), it has recently been adapted to other languages to provide research resources, and cross-lingual

CDI databases are now available. The next sub-section describes the databases we used.

2.1 American, Spanish and Danish CDI databases

We use American (Dale et al., 1996), Mexican Spanish (Dale et al., 1996), Danish (Madsen, 2008) database in the Lex 2005 CDI database (Jørgensen, et al., 2010)

American inventory for WG has 396 words. The comprehension and production norms are calculated from inventory result for children from 8 to 16 month ages. American inventory for WS has 680 words for children from 16 to 30 month ages. Mexican Spanish inventory for WS has 427 words. The comprehension and production norms are calculated from inventory result for children from 8 to 18 month ages. Mexican Spanish inventory for WS has 681 words for children from 16 to 30 month ages.

2.2 Japanese CDI database

We collected cross-sectional data from 1,852 mothers living in Nara, Osaka, and Kyoto with 10~32 month-old children, and these women performed the Japanese version of CDI at our laboratories over about six years from April, 2006. The Japanese inventory for WG has 448 words for 10 to 22 months. WS has 711words for 20 to 32 month-old children..

Table 1. Number of children in WS and WG for each language.

Language	Number of children	
	WS	WG
American	1461	1089
Mexican Spanish	778	1094
Danish	3714	2398
Japanese	1506	346
Croatian	250	377
French (Quebec)	827	537
Italian	753	648
Korean	156	40
Latvian	500	183
Norwegian	9304	2926
Slovak	1065	657
Turkish	2422	1115

2.3 Croatian, French_(Quebec), Italian, Korean, Latvian, Norwegian, Slovak and Turkish databases

CDI has also been applied to other languages. The following databases were extracted from Wordbank's available databases (Frank et al., 2016): Croatian (Kovacevic, Babic, Brozovic, 1996), French (Quebec) (Trudeau, Sutton, 2011), Italian (Caselli, Casadio, Bates, 1999), Korean, Latvian, and Norwegian (Simonsen, Kristoffersen, Bleses, Wehberg, Jørgensen, 2014), and Slovak and Turkish (Ay, 2009). The data were downloaded on 9/17/17.2.4 Number of children for each language database

The number of children for each language is shown in Table 1.

3. Fitting the acquisition curves by logistic functions

By monthly categorizing the obtained CDI data, we calculated the acquisition rates of children who comprehend and produce a particular word every month. Here the rates, which were calculated from insufficient data, were treated as missing values. Then we modeled these curves using logistic functions with respect to age in days:

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

$$f'(x) = \frac{a'e^{c'x+b'}}{1+e^{c'x+b'}}, \quad (2)$$

where f and f' are the logistic functions for word comprehension and word production. First, Eq. (1) is calculated by the nonlinear least mean square method to fit the rates of infants who produce a word. Here we introduce a constraint where the probability must be one or less. To satisfy this constraint, if a is bigger than 1.0, it is set to 1.0, and then b and c are recalculated by the nonlinear least mean square method.

Next we calculated Eq. 2. However, we found that for some words, the comprehension acquisition rates were not appropriately calculated since we didn't have as much data for these words as for the word production data. This lack of data was caused by the difficulty that mothers verify the their children's vocabulary after their children comprehend a large number of words. Thus we set constraint $f(x) \leq f'(x)$ under which a', b', and c' should be obtained. However, simultaneously obtaining a', b', c', a, b, and c under the constraint is mathematically difficult. Therefore instead of that complicated constraint, we introduce a simple constraint, $a \leq a'$, in the following optimizing method:

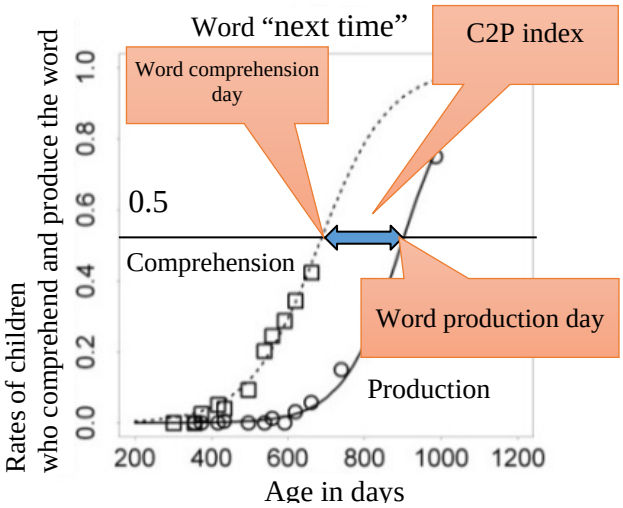

Fig. 1 Logistic functions fitting production acquisition and comprehension rates of kondo (next time) with constraint and calculating C2P index .

To estimate correctly the production rate, we introduce a constraint where the word comprehension rate should exceed the word production rate. a', b' and c' are obtained by the non-linear least mean square method ; if a>a', a' is fixed to a, and then b' and c' are recalculated by non linear least mean square method. If a' > 1.0, a' is fixed to 1.0, and then b and c are recalculated.

Fig. 1 shows an example of estimated logistic functions for the acquisition rates of kondo (next time). These approximations work well for the acquisition and word production rates. An example of this situation is shown in Fig. 1, where the dotted line is the comprehension rate curve. The solid line in Fig. 1 shows the obtained production rate.

4. Word comprehension-to-production index

The ratio of the parts of speech produced within a certain period sequence was previously used to investigate which part of speech predominates. One problem with such ratios is that since they can only be obtained for parts of speech, they cannot determine which word is difficult by a word-by-word examination. Because of this, the rate is strongly affected by such environments as culture and parent input

To evaluate word difficulty in such a word-by-word fashion, we used the word comprehension day, the word production day, and the period between them. We call this period the comprehension-to-production (C2P) index. Although some might think that it only evaluates a partial learning process of children, it is important to subdivide the learning process and investigate each part of it to precisely understand the entire learning process. To calculate this index, we first define the word comprehension and production days as when 50% of the children respectively comprehend and produce a particular word. These days were determined by approximating the word comprehension and production rate curves by two logistic functions, setting the functions to 0.5, and solving them by the Newton method. Fig. 1 shows an example of the C2P index for kondo (next time).

5. Investigation of noun predominance for word acquisition days and periods

We used the American database to investigate which indexes are good measures to evaluate noun predominance: the comprehension days, the production days, or the C2P indexes. To classify the words into nouns and verbs, we used Caselli's part-of-speech classification (Caselli et al.) and calculated the comprehension days, the production days, and the C2P indexes for the words except those whose C2P indexes couldn't be calculated.

Fig. 2 shows the noun and verb distributions for the comprehension days for the American database. The average number of comprehension days for verbs was 480, and for nouns it was 498, showing a small difference in the number of days between them ($p < 0.2$). In terms of the comprehension days, nouns do not predominate verbs. Fig. 3 shows the noun and verb distributions for the American database for the word production days. The average for verbs was 731 days, and for nouns it was 681 days, showing significant differences in the number of days ($p < 0.001$). Nouns predominate verbs in terms of word production days.

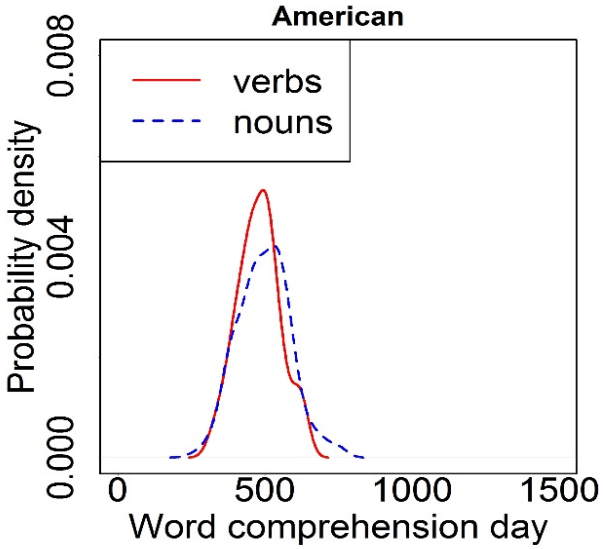

Fig. 2 American word distributions of nouns vs. verbs with respect to word comprehension days .

Considering the process of word acquisition, the word production day can be divided into two processes: the comprehension process and the process for the C2P index. From Fig. 2 we confirmed that word comprehension days do not contribute to noun predominance, suggesting that the C2P index period primarily contributes to noun predominance.

Fig. 4 shows the noun and verb distributions for the American database for the C2P index. The average C2P index of verbs was 251 days, and for nouns it was 183 days, showing significant differences in the number of days ($p < 0.001$). This result shows that in the periods between the comprehension and production days, nouns

strongly predominate verbs. The process from word comprehension to word production strongly affects noun predominance in word-learning.

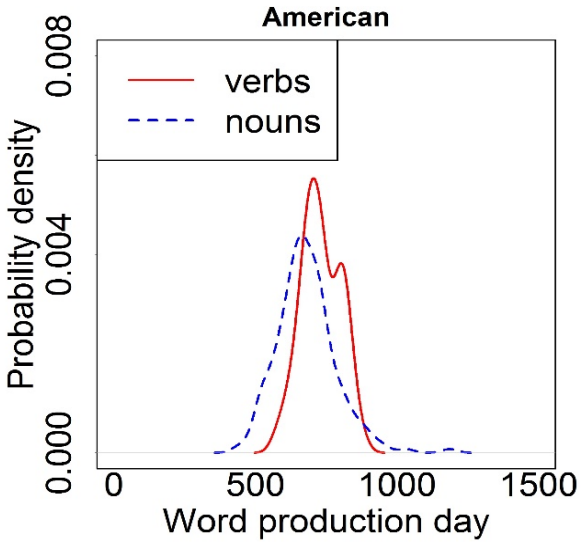

Fig. 3 American word distributions of nouns vs. verbs with respect to word production days .

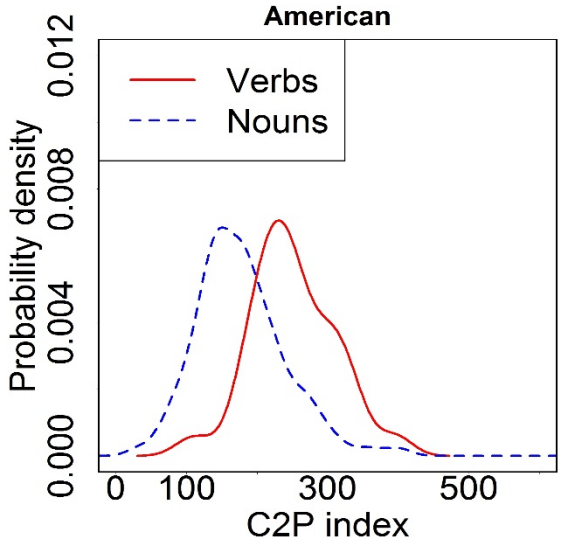

Fig. 4 American word distributions of nouns vs. verbs with respect to C2P index .

Fig. 5, Fig. 6 and Fig.7 show the noun and verb distributions for the Danish database for the comprehension day, the production day, and the C2P index. These results resemble those of the American database and also show that the process from word comprehension to word production strongly affects noun predominance in word-learning.

We calculated the comprehension days, the production days, and the C2P indexes for all of these languages to investigate the generality of this result using the cross-lingual database described in Section 2. Tables 1, 2, and 3 show the comprehension days, the production days, and the C2P indexes of nouns and verbs for the target languages. We also evaluated the value differences between nouns and verbs using a t-test. The results show noun predominance in the C2P indexes among all the languages except for Slovak and Turkish.

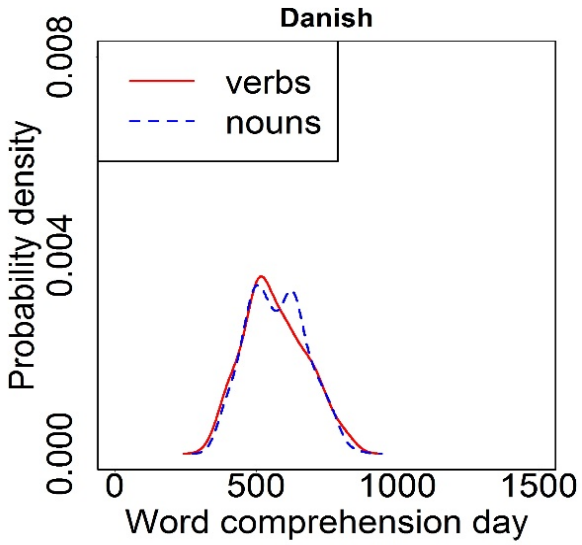

Fig. 5 Danish word distributions of nouns vs. verbs with respect to word comprehension days .

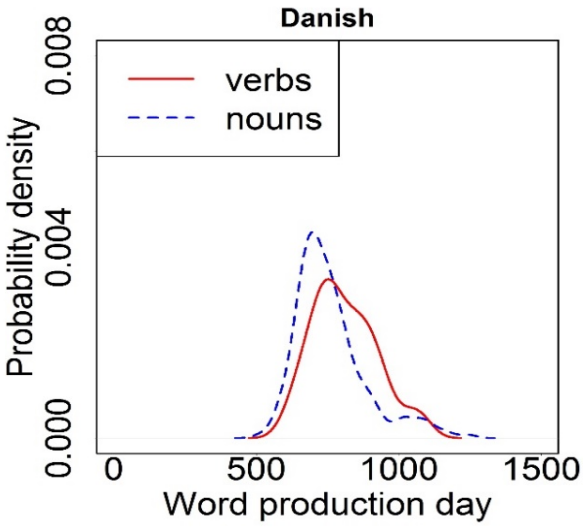

Fig. 6 Danish word distributions of nouns vs. verbs with respect to word production days .

This result suggests that the process between a word's comprehension and its production is a significant factor of predominance in noun learning by children and strongly supports the Gentner relational relativity hypothesis, because, to produce a verb, the surrounding words must be understood. The C2P index is a good measure to evaluate this factor because we confirmed that it evaluates the word acquisition difficulty of individual verbs and indicates that investigation of the C2P index of verbs might reveal the infant acquisition process of syntax.

Table 4 shows that C2P is also a good measure to examine language characteristics.

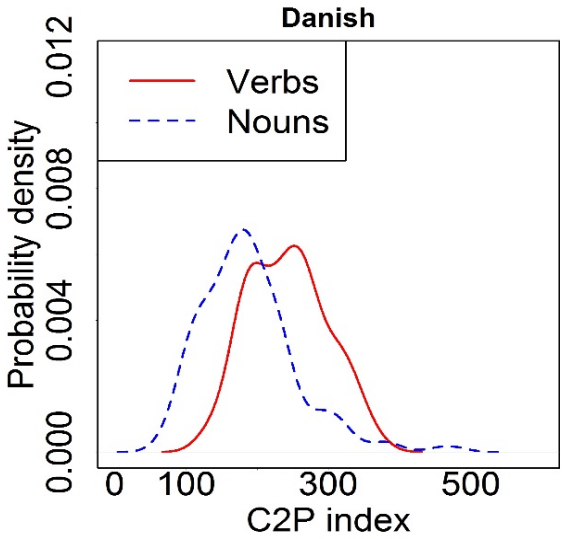

Fig. 7 Danish word distributions of nouns vs. verbs with respect to C2P index

6. Conclusion

This paper defines a comprehension-to-production (C2P) index that measures the difficulty of a partial word learning process from comprehension to production to investigate whether nouns have pre

dominance over verbs in word learning by children. We evaluated C2P indexes for cross-languages using cross-lingual CDI databases and confirmed that they indicate noun predominance in word learning and that the process between a word's comprehension and its production is a significant factor of predominance in noun learning by children. The experiment in this paper strongly supports the Gentner relational relativity hypothesis and argues that C2P is also a good measure to examine verb and language characteristics.

Table 2. Word comprehension days of nouns and verbs for target languages and t-test results.

Language	Word comprehension days		P-value
	Nouns	Verbs	
American	498	480	0.13
Danish	572	568	0.79
Mexican Spanish	509	483	0.12
Japanese	584	557	0.009
Croatian	439	453	0.35
French_(Quebec)	461	507	0.047
Italian	539	469	0.009
Korean	480	482	0.94
Latvian	425	431	0.83
Norwegian	497	528	0.24
Slovak	416	397	0.46
Turkish	468	469	0.95

Table 3. Word production days of nouns and verbs for target languages and t-test results.

Language	Word production days		P-value
	Nouns	Verbs	
American	681	731	4.8e-05
Danish	760	810	0.007
Mexican Spanish	736	817	4.0e-07
Japanese	773	806	0.002
Croatian	694	760	0.001
French_(Quebec)	659	770	6.0e-05
Italian	739	775	0.056
Korean	699	756	0.01
Latvian	697	775	0.003
Norwegian	688	756	0.018
Slovak	708	682	0.33
Turkish	744	761	0.53

Table 4. C2P indexes of nouns and verbs for target languages and t-test results.

Language	C2P index (days)		P-value
	Nouns	Verbs	
American	183	251	6.0e-10
Danish	188	242	2.8e-08
Mexican Spanish	227	334	4.0e-16
Japanese	188	249	5.7-11
Croatian	255	307	8.5e-05
French_(Quebec)	198	263	3.9e-06
Italian	200	306	4.0e-10
Korean	218	273	0.0024
Latvian	272	345	2.2e-05
Norwegian	191	227	0.002
Slovak	291	284	0.70
Turkish	276	291	0.18

7. Acknowledgements

A part of this study was supported by a Grant-in-Aid for Scientific Research (B) 17H02190.

8. Bibliographical References

Benedict, H. (1979). Early lexical development: Comprehension and production. Journal of child language, 6(2), 183-200.

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

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. Language, 2(1), 301-334.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. Language acquisition and conceptual development, 3, 215-256.

Maguire, M. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). A Unified Theory ofWord Learning: Putting Verb Acquisition in Context. In Action meets word: How children learn verbs, 364-391.

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

9. Language Resource References

Ay, S. (2009). Essays on Turkish linguistics: proceedings of the 14th International Conference on Turkish Linguistics, August 6-8, 2008 (Vol. 79): Otto Harrassowitz Verlag.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., et al. (2008). The Danish Communicative Developmental Inventories: validity and main developmental trends. Journal of child language, 35(3), 651-669.

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(01), 69-111.

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

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.

Jørgensen, R. N., Dale, P. S., Bleses, D., & Fenson, L. (2010). CLEX: A cross-linguistic lexical norms database*. Journal of Child Language, 37(2), 419.

Kovacevic, M., Babic, Z., & Brozovic, B. (1996). A Croatian language parent report study: Lexical and grammatical development. Paper presented at the Seventh International Congress for the Study of Child Language, Istanbul, Turkey.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23.

Trudeau, N., & Sutton, A. (2011). Expressive vocabulary and early grammar of 16-to 30-month-old children acquiring Quebec French. First language, 31(4), 480-507.