## Title
Symbol grounding and system construction in the color lexicon



Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors
Imai, Mutsumi
Hidaka, Shohei
Saji, Noburo
et al.

Publication Date
2018

# Symbol grounding and system construction in the color lexicon

**Mutsumi Imai (imai@sfc.keio.ac.jp)**
Faculty of Environment and Information Studies, Keio University, 5322 Endo,
Fujisawa, Kanagawa 252-0882 Japan

**Shohei Hidaka (shhidaka@jaist.ac.jp)**
School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan

**Noburo Saji (saji@kamakura-u.ac.jp)**
Department of Child Psychology, Kamakura Women's University, 6-1-3 Ofuna,
Kamakura, Kanagawa 257-0056 Japan

**Masato Ohba (ohbam@sfc.keio.ac.jp)**
Faculty of Environment and Information Studies, Keio University, 5322 Endo,
Fujisawa, Kanagawa 252-0882 Japan

**Abstract**

This research investigated the acquisition process of the color lexicon, specifically how color words are initially grounded and develop into the lexical system possessed by the adults in the ambient language. We conducted a longitudinal study in which Japanese-learning 2-year-olds were tested every month on their understanding of basic words denoting 8 chromatic colors, continuing until they were able to map these words onto their referents consistently. The results strongly endorse the view that acquisition of the color lexicon should be characterized as a process of system construction, through which children reorganize prelinguistic color categories onto the linguistic categories of the ambient language, thereby representations of individual words are continuously refined along with the refinement of the representation of the system as a whole.

**Keywords:** color word learning; lexical acquisition; symbol grounding; Bayesian model

## Introduction

A great number of researchers have investigated how children fast map a novel word to its referent and generalize this mapping to other potential referents. A majority of the studies have been conducted under the implicit assumption that the inference of the meaning of a word takes place in isolation from other words in the lexicon, especially in the earliest stages of lexical development. However, the lexicon is not merely an assembly of words each standing on its own; it is a complexly structured system, in which words are contrasted to one another along multiple dimensions at multiple levels (cf. Saussure, 1916/1977). Thus, simply accumulating a number of words that are not integrated in the lexicon does not lead to the acquisition of the complex-structured lexicon possessed by adult native speakers (Clark, 1987; Haryu & Imai, 2002; Saji et al., 2011).

Take color words, for example. We cannot say that a child has acquired the meaning of the English word *red* if all he knows is that red is the color of apples: the child has to learn the range of colors that can be named by this word in the English-speaking community. Children seem to perceive the color spectrum categorically well before they start to show the sign of understanding color words (Bornstein, 1985b; Skelton et al., 2017; Yang et al., 2016). However, because languages differ widely in the ways in which they divide the continuous visible spectrum of color by color names (e.g., Berlin & Kay, 1969; Cook, et al., 2005; Roberson et al., 2005), to be able to apply the word '*red*' to its appropriate range of referents, children need to know how *red* contrasts with *pink*, *orange, purple,* and where the boundaries are delineated between these words in the lexicon of adult English speakers.

Importantly, researchers have long noted that color word acquisition is notoriously slow (Bornstein, 1985b; Kowalski & Zimiles, 2006; Sandhofer & Smith, 1999). Bornstein (1985a) suggested that color words were not reliably acquired until around 4 to 7 years of age. Some developmental researchers have proposed that it is difficult because the concept of color is difficult to grasp for young children: to understand that the term refers to a property, children need to extract the perceptual property of 'color' from diverse kinds of objects that differ not only in color but also in other properties such as shape, size, and texture (e.g., Bornstein, 1985a; Kowalski & Zimiles, 2006; Sandhofer & Smith, 1999). Wagner, Dobkins, and Barner (2013), in contrast, argued that the difficulty lies in discovering the language specific boundaries of color categories, as such boundaries are not marked anywhere in the visible environment (see also Braisby & Dockrell, 1999).

Considering these issues, it is not likely that color word acquisition amounts to a collection of simple mapping process of prelinguistic color categories onto corresponding word forms (cf., Yang et al, 2016); it must involve, at least in part, a process through which children (re)organize prelinguistic color categories onto linguistic categories of the ambient language. To investigate the detailed process of color lexicon acquisition as an entire lexical system, Saji, Asano, Oishi and Imai (2015, in submission) examined how three- to five-year-old children named 93 color patches that covered the entire spectrum of the color space, including both typical and less typical color swatches denoted by basic color words in Japanese. Largely supporting Wagner et al.'s view (2013), Saji et al. found that, by 3-year-olds, children are able to apply most basic words (*aka* (red), *ao* (blue), *ki* (yellow), *midori* (green), *pinku* (pink), *cha* (brown), *orenji* (orange), *murasaki* (purple)) to their typical referents, but that the boundaries were very messy at that age; the boundaries became delineated more cleanly with development, but even at 5, the children's categories were still not like those of adults.

These results indicate that children's color lexicon develops gradually over time as a *connected system*, in which the elements, i.e., the color words, are related one another to constitute a meaningful structure. However, as Saji et al. studied the growth of the color lexical system from 3-years onward, little is known about how the system is started, expanded, and structured before that time. For example, in Saji et al. (2015), the proportion of correct naming for the typical referents of basic color names varied across words. Typical colors for '*aka*' (red), '*ao*' (blue) and '*ki*' (yellow) were named most accurately, while '*cha*' (brown) and '*murasaki*' (purple) were applied much less accurately. Does this pattern reflect the order in which these words are grounded onto children's color lexicon? More importantly, what the representation of the color lexicon as a whole is like at the beginning stage of acquisition and how does it change developmentally?

To address these issues, in the present research, we conducted a longitudinal study in which Japanese-learning 2-year-olds were tested every month starting from 29.8 months-old (on average)-- the time in which a majority of children know few color words-- on her understanding of basic words denoting 8 chromatic colors ('*aka*' (red), '*ao*' (blue), '*ki*' (yellow), '*midori*' (green), '*pinku*' (pink), '*cha*' (brown), '*orenji*' (orange), '*murasaki*' (purple)), and continuing until they were able to map all of these words onto their referents consistently.

## Method

**Participants** A total of 49 Japanese-learning children (31 boys) participated. Testing was started when each child was between 24 and 35 months (Mean = 29.83). Each child was tested every month until s/he was able to map all of the 8 words to their correct referent colors in two consecutive months. Eight additional children who succeeded in all eight of the word-referent mappings on the first test session were not tested further. Data from a total of 413 sessions were obtained from 49 children. The mean number of sessions per children were 9.45, which means that it took children 9.45 months on average to establish the mappings between the 8 basic chromatic words and their corresponding colors (SD = 5.65, range = 1-18).

**Stimuli.** A picture book consisted of four stories, each containing objects (e.g., balloons) of 8 basic chromatic colors ('*aka*' (red), '*ao*' (blue), '*ki*' (yellow), '*midori*' (green), '*pinku*' (pink), '*cha*' (brown), '*orenji*' (orange), '*murasaki*' (purple)) were prepared. In each story, the experimenter asked the child to name two colors. The story-color pairings were counter-balanced. The color tokens were chosen based on a pre-study with adult Japanese speakers. Twelve adult speakers of Japanese were asked to name 93 color swatches from the Practical Color Co-ordinate System (PCCS) developed by Japan Color Research Institute[1]. The color swatch that received the highest agreement by the adults was chosen to be the referent of each word (see Saji et al., 2015, in submission).

**Procedure.** In each session, children were tested individually in a room at the preschool/daycare they were enrolled in. The experimenter asked the child to point to the object of the color indicated in the story. The same stimuli and procedure were used throughout the testing sessions. The lighting condition was controlled by using Solax XC-100AF (Seric Ltd.), which simulated natural daylight (D55).

## Results and Discussion

### Analysis I: Individual trajectories of acquisition.

We first visualized the developmental trajectory for each child (Figure 1). To adjust the variability in performance across sessions within individuals, we calculated the *moving average of accuracy* for each child, by taking the average of the response accuracy across the moving window of five consecutive sessions, fixing the third as the target. Each line shows how the accuracy for each child changes through development. There was a large individual difference in the speed of development. The overall accuracy increased

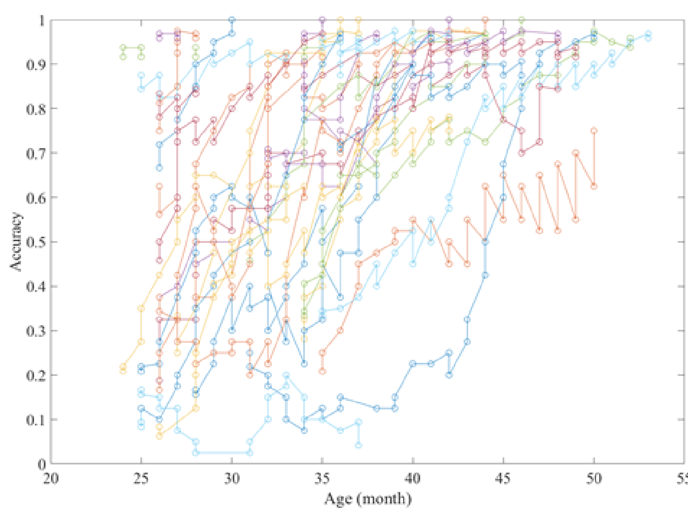



Figure 1: The moving average of accuracy rates for each child as a function of age.

[1] PCCS consist of 14 'tone' categories, each of which has 24 hues. Tone is a compound concept of metric lightness and metric chroma (Nayatani, 2003). Colors with the same PCCS tone have the same perceived lightness and chroma irrespective of hue.

steadily with age, but when seen more finely, a number of children underwent minor regression, suggesting that there is continuous restructuring of the lexicon.

## Analysis II: Learning speed for mapping each word to the corresponding color

We first estimated the learning rates for each of the 8 colors by a logistic regression model separately conducted for each color with the age (in months) as fixed effects. Here, we considered *accuracy* as the degree in which a given word is correctly mapped to its corresponding color, without considering the degree in which the word was overextended to wrong colors. Because the learning speed is expected to vary largely across children, the age factor was included in the model as a by-subject random slope factor. The estimated month-by-month trajectories of the correct response rates for the 8 words is shown in Figure 2. Four words—*aka* (‘red’), *ao* (‘blue’), *ki* (‘yellow’), and *pinku* (‘pink’) — showed relatively sharp growth rates, as the probabilities passed 0.8 around 30 months of age. In contrast, the developmental slopes for the remaining four words—*orenji* (‘orange’), *midori* (‘green’), *cha* (‘brown’), *murasaki* (‘purple’) —appears more gradual and reached 80% accuracy only after 35 months. The results are somewhat different from the pattern predicted by Berlin and Kay’s universal hierarchy hypothesis in two respects: *midori* (‘green’), which is supposed to be among the earliest acquired chromatic colors, was learned slowly while *pinku* (‘pink’) was learned fast. However, this pattern is consistent with Saji et al. (2015)’s results from a cross-sectional study.

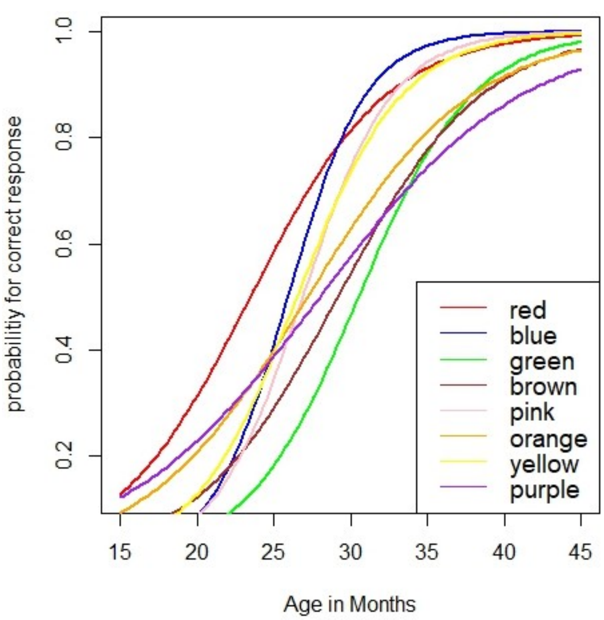



Figure 2: The logistic regression models predicting the trajectory of learning rates for each of the 8 words

with older children (3-5 year olds), suggesting that cultural effects may appear in learning of color words even at this earliest stage.

## Analysis III: Developmental trajectory of the color lexical system.

Analysis II examined the learning rate of each color word across children. This analysis, however, does not reveal how precisely children understand the meaning of each color word, or how the color lexicon develops as a connected whole. in that it does not consider the degree of overextension for each word (e.g., using ‘*aka*’ to name the blue color). To understand the development of the color lexical as a connected system, we should see the pattern of over- and under-extension across the 8 words, and how this pattern changes as the learning of the color lexicon evolves toward more matured system that is closer to that possessed by adults.

For this goal, we attempted to characterize the developmental changes in the similarity structure underlying the 8 words quantitatively. Because the similarity structure of the color lexicon is expected to be fluid and continuously changing in the course of development, we employed a Bayesian version of the Similarity Choice Model (SCM, Townsend & Landon, 1982).

SCM defines the conditional probability to choose a color token (e.g., ‘yellow’) for a target word (e.g., *aka*), as a function of two parameters: the *similarity* parameter and the *bias* parameter. The former represents the degree of correspondence between a word and a color token (e.g., ‘*ki*’ and red, *‘aka*’ and yellow, etc), where each pair of the corresponding color word and token (e.g., ‘*aka*’ (‘red’) and red) is assumed to have a fixed similarity 0, and to be interchangeable (i.e., the similarity parameters between *ki* and ‘red’ and that between *aka* and ‘yellow’ are the same). The latter parameter represents the degree of a general response bias for choosing a particular color token (e.g., choosing the token of ‘red’ whichever word was given as the target).

Mathematically, the similarity parameter and the bias parameter are estimated as the posterior distribution $P\left(\eta_{ij},\beta_j\right) \propto \prod_{i,j} P(i|j)^{n_{ij}}$ of $\eta_{ij}$ and $\beta_j$ given the choice count $n_{ij}$ , where the conditional probability $P(j|i)$ is the probability to choose the *j*th color token given the *i*th color word, following the Equation (1) of the Similarity Choice Model (Townsend and Landon, 1982) for all $1 \le i,j \le 8$.

$$P(j|i) = \frac{e^{\eta_{ij}+\beta_j}}{\sum_{j=1}^{8} e^{\eta_{ij}+\beta_j}} \qquad (1)$$

Five hundred thousand samples of the similarity and the bias parameters were obtained by the Metropolis-Hasting method. However, the samples prior to the one with the maximum posterior probability was discarded as the transient. After this operation, approximately 100,000 samples were retained for further analyses.

For each sample, we conducted the multidimensional scaling on the corresponding $8 \times 8$ dissimilarity matrix with its $(i, j)$ element $-n_{ij}$, using the function 'mdscale' (with the default option) in the MATLAB statistical toolbox to visualize the similarity structure[2].

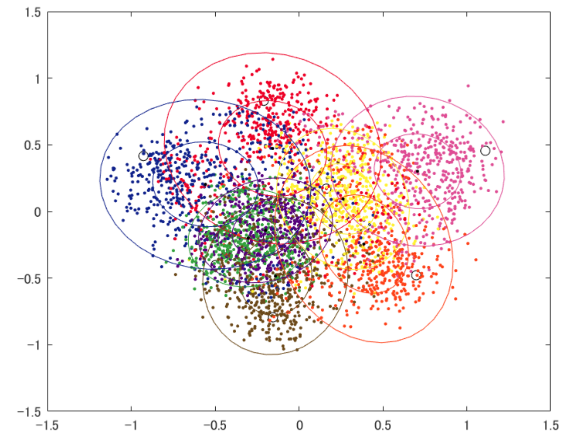
To capture the developmental changes of the similarity structure more closely, we divided the dataset into the low-, middle- and high-performing periods, and conducted SCM separately for each period. The sessions whose *moving average* of accuracy (see Analysis I) were lower than 50% were considered to belong to the low-performing period and those whose moving accuracy exceeded 75% were considered as the high-performing period; 114, 98, and 201 sessions were included in the low-, middle-, and high-performing periods, respectively.

Figure 3 (a)-(c) show the 2 dimensional plane of MDS for the three performing periods. Each point on the MDS plane represents the coordinate of the corresponding color for each estimated similarity parameters. The two ellipses represent the equal-variance boundary for 1-SD and 2-SD. Note that the size of the ellipsis reflects the amount of variability in the estimated similarity between two words in some degree; but the ellipsis size should not be taken as the direct index of inconsistency of word-color correspondence, because it is influenced by the Procrustes transformation, which is a method of mathematical transformation for preserving the geometric 'shape' of an object in overlaying the MDS coordinates from different samples onto a single plane. By this operation, the size of an ellipsis tends to become larger as it goes away from the center of the plane. In each MDS configuration, the distance between two colors directly indicates the likelihood of confusion by children rather than degree of overlapping visually observed in the figure, because observed overlaps on the averaged MDS plane may have arisen from seemingly close points of multiple distinct MDS configurations, which are not directly comparable.

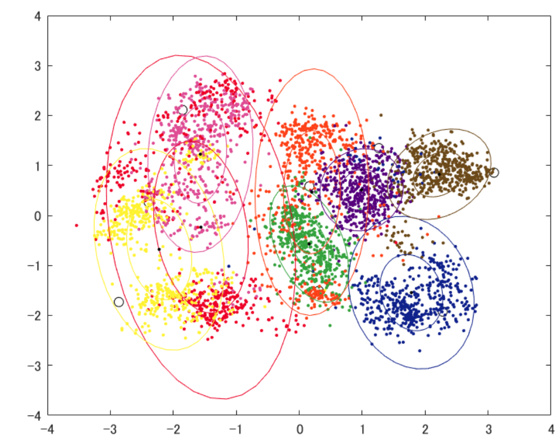
For example, in Figure 3b, 'purple' and 'green' are closer on average than 'red' and 'pink.' This indicates that green and purple were confused each other in a greater degree than red and pink were.

[2] As SCM assumes that the similarity parameter between a particular pair of a word (e.g., *aka*) and a token (e.g., red) is symmetric, each point in the MDS plane represents both a word and the corresponding token simultaneously.

(a)




(b)




(c)

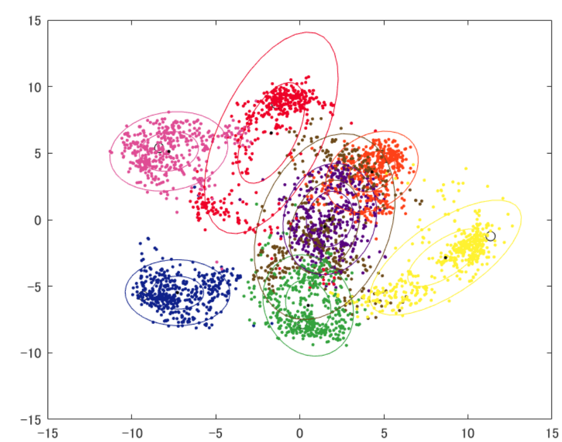



Figure 3 a-c: The results of the multidimensional scaling. The points on the MDS plane show the (dis)similarity parameters calculated from each of the randomly sampled 500,000 sets, which were drawn from the posterior distribution and 2-SD. (a): low-performing period: (b): middle-performance period: (c): high-performing period.

When Figure 3 a-c are compared, it is clear that the configuration of the similarity among the 8 basic color words goes through striking changes across the three periods. At the earliest period, all of the 8 colors are largely lumped together. In the low-performance period (Fig 3a), the 8 colors were concentrated around the center, indicating a great degree of confusion in the system as a whole. In particular, *midori* ('green'), *murasaki* ('purple'), and *cha* (brown) are almost indistinguishable from one another, suggesting a great degree of confusion among them and that the similarity structure in children's color lexicon is largely different from the similarity space organized by adults (Saji et al., 2015; Kuriki et al., 2017).

By the time children have reached approximately 50% accuracy in word-referent mappings, a cluster of the warm color words has been formed. *'Ao*' (blue) is clearly distinguished from other colors but the four late grounded words ('*orenji*, *midori*, *murasaki*, *cha*') are still confused with one another. In Figure 3c, the distance between every pair of words became greater, suggesting that the words were much less confused. At this stage, '*pinku*, *ao*, *aka, ki*' were located away from the center toward the edge of the MDS plane, while *'orenji*, *murasaki*, *ch*a, *midori*' were still close to one another around the center. This pattern indicates that the former four words were no longer overextended at this stage but the latter four words were still confused one another to some extent.

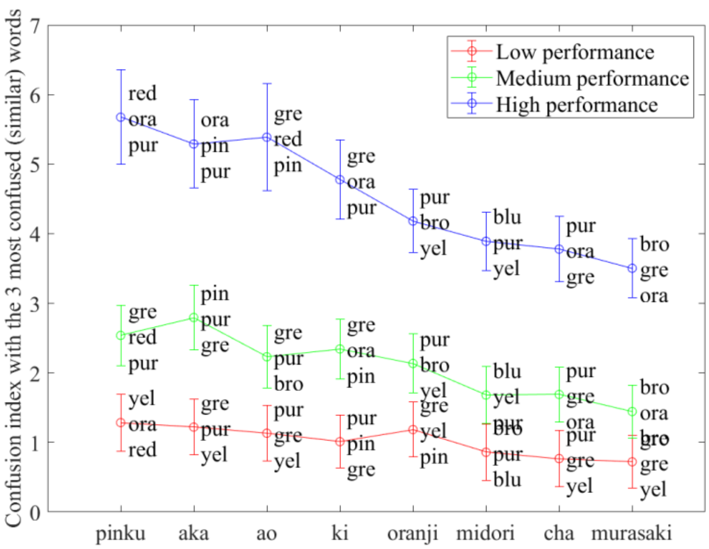



Figure4. The degree of confusion (overextension) for each of the 8 colors. The values on the Y axis represents the degree of confusion (overextension), which was obtained by calculating the average of the dissimilarity value between the target color and each of the three most overextended colors, separately in each of the three performance groups (Low, Med, High). Larger values on the Y axis show high dissimilarity (i.e., a small degree of confusion) between the word and non-target colors. At each point, the three most overextended colors (first three letter) for the word are shown in the descending order.

In order to quantify the degree of confusion (overextension) for the eight color words, the dissimilarity matrix was calculated for the samples drawn from the Bayesian similarity choice model. For each word in each of the three performance group, the three most confused colors (i.e. those with least dissimilar parameters $\eta_{ij}$ ) were identified, and the index for overextension was calculated by averaging the dissimilarity values between the target word and the three most confused colors (Figure 4). The color words along the X-axis was arranged, from left to right, from the least confused to the most confused word. This analysis revealed that that '*cha*, *murasaki*, and *midori*' were highly confused, while '*pinku*, *aka*, and *ao*' were least overextended. To test whether the degree of overextension differed across the eight colors, and whether the effect was modified by the performance groups, we carried out a two-way ANOVA on the 276 samples of dissimilarity drawn from the Gibbs sampler, where the number of samples is equalized to the original sample size (828 sessions in total, 276 on average for each group) in order to have the equal level of statistical power. The analysis found a significant main effect for the performance group ($F(2, 6600)= 7241.4$, $p < 0.01$) and the color ($F(7, 6600)=3226.2$, $p < 0.01$), but the main effects were modified by a significant interaction between them ($F(14, 6600)= 33.3$, $p < 0.01$).

## General Discussion and Conclusion

This research investigated the acquisition process of the color lexicon, asking how color words are initially grounded and develop into the lexical system through a longitudinal data collection. When the testing began, a majority of children seemed to understand that the words are the names of colors, and that color is a property separate from other visual properties such as shape, pattern and texture. However, the learning of mappings between the words and their typical referents—just 8 of them—is slow, requiring more than 9 months on average and continuous reorganization of the boundaries of these words.

We identified that the 8 color words are acquired in the order of *pinku*, *aka*, *ao*, *ki, orenji*, *midori*, *cha*, and *murasaki* both in light of the accuracy of mapping to the category center, and in light of correct category boundaries. This order is somewhat different from the pattern predicted by Berlin and Kay's universal hierarchy hypothesis in two respects: *midori* ("green"), which is supposed to be among the earliest acquired chromatic colors, was learned slowly while *pinku* ('pink') was learned fast. However, the contribution of this research is that it showed how the early color lexicon was represented and evolved as a connected system quantitatively.

This research leaves some unsolved questions. In particular, it is still not clear how biological and cultural factors play a role in constructing the initial color lexical system. Prior to this study, we had hypothesized that biological factors would play the primary role for boosting the system, and the contribution of cultural factors (e.g., cultural specific preferences for particular colors or cultural specific distributions of color words) would come in only at later stages. However, the results of the present research suggest that this expectation was too simplistic. The pattern of the results seems to show that both factors are jointly operating from the very beginning. Further research is needed to follow up on this possibility.

In any case, what has been made clear is that children are not acquiring color words by merely attaching labels to perceptual categories (cf. Yang et al., 2016). The present research strongly endorses the view that acquisition of the color lexicon should be characterized as a process of system construction, through which children (re)organize prelinguistic color categories onto linguistic categories of the ambient language.


## Acknowledgments

We thank Yukie Ujihara, Kanako Yasufuku, Sakiko Hayashida for data collection. This research was supported by MEXT KAKENHI to Imai (Grant-in-Aid for challenging Exploratory Research #16K13224)


## References


Berlin, B., & Kay, P. (1969). *Basic color terms: Their university and evolution*. California UP.

Bornstein, M. H. (1985a). Colour-name versus shape-name learning in young children. *Journal of Child Language*, *12*(2), 387-393.

Bornstein, M. H. (1985b). On the development of color naming in young children: Data and theory. *Brain and language*, *26*(1), 72-93.

Braisby, N., & Dockrell, J. (1999). Why is colour naming difficult ? *Journal of Child Language*, *26*(1), 23-47.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1-33). Hillsdale, NJ: Erlbaum.

Cook, R. S., Kay, P. & Regier, T. (2005). The world color survey database: History and use. In Cohen, H. & Lefebvre, C. (Ed.), *Handbook of Categorisation in the Cognitive Sciences*. Amsterdam: Elsevier.

De Saussure, F. (1916). *Cours de linguistique générale* (C. Bally & A. Sechehaye, Ed.), trans. W. Baskin, Course in General Linguistics, Glasgow: Fontana/Collins, 1977.

Franklin, A., & Davies, I., (2004). New evidence for infant colour categories. *British Journal of Developmental Psychology*, *22*(3), 349-377.

Haryu, E., & Imai, M. (2002). Reorganizing the lexicon by learning a new word: Japanese children's interpretation of the meaning of a new word for a familiar artifact. *Child development*, *73*(5), 1378-1391.

Imai, M., & Haryu, E. (2001). Learning proper nouns and common nouns without clues from syntax. *Child development*, *72*(3), 787-802.

Kowalski, K., & Zimiles, H. (2006). The relation between children's conceptual functioning with color and color term acquisition. *Journal of Experimental Child Psychology*, *94*(4), 301-321.

Kuriki, I., Lange, R., Muto, Y., Brown, A. M., Fukuda, K., Tokunaga, R., Lindsey, D., Uchikawa, K., & Shioiri, S. (2017). The modern Japanese color lexicon. *Journal of vision*, *17*(3), 1-18.

Roberson, D., Davidoff, J., Davies, I.R.L., & Shapiro, L.R. (2004). The development of color categories in two languages: A longitudinal study, *Journal of experimental Psychology: General*, *133*(4), 554-571.

Saji, N., Imai, M., Saalbach, H., Zhang, Y., Shu, H., & Okada, H. (2011). Word learning does not end at fast-mapping: Evolution of verb meanings through reorganization of an entire semantic domain. *Cognition*, *118*(1), 45-61.

Saji, N., Asano, M., Oishi, M., & Imai, M. (2015). How do children construct the color lexicon? : Restructuring the domain as a connected system. *Proceedings of the 37th Annual meeting of Cognitive Science*. 2080-2085.

Sandhofer, C. M., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. *Developmental Psychology*, *35*(3), 668.

Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences*, *114*(21), 5545-5550.

Nayatani, Y. (2003). Adequateness of a newly modified opponent-colors theory. *Color Research & Application*, *28*, 298-307.

Townsend, J. T., & Landon, D. E. (1982). An experimental and theoretical investigation of the constant-ratio rule and other models of visual letter confusion. *Journal of Mathematical Psychology*, *25*(2), 119-162.

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition*, *127*(3), 307-317.

Yang, J., Kanazawa, S., Yamaguchi, M. K., & Kuriki, I. (2016). Cortical response to categorical color perception in infants investigated by near-infrared spectroscopy. *Proceedings of the National Academy of Sciences*, *113*(9), 2370-2375.