A Computational Model for Child Inferences of Word Meanings via Syntactic Categories for Different Ages and Languages

Yuji Kawai, Yuji Oshima, Yuki Sasamoto, Yukie Nagai, *Member, IEEE*, and Minoru Asada, *Fellow, IEEE*

***Abstract*—Children exploit their morphosyntactic knowledge in order to infer the meanings of words. A recent behavioral study has reported developmental changes in word learning from three to five years of age, with respect to a child's native language. To understand the computational basis of this phenomenon, we propose a model based on a hidden Markov model (HMM). The HMM acquires syntactic categories of given words as its hidden states, which are associated with observed features. Then, the model infers the syntactic category of a new word, which facilitates the selection of an appropriate visual feature. We hypothesize that using this model with different numbers of categories can replicate the manner in which children of different ages learn words. We perform simulation experiments in three native language environments (English, Japanese, and Chinese), which demonstrate that the model produces similar performances as the children in each environment. Allowing a larger number of categories means that the model can acquire a sufficient number of obvious categories, which results in the successful inference of visual features for novel words. In addition, cross-linguistic differences originating from the acquisition of language-specific syntactic categories are identified, i.e., the syntactic categories learned from English and Chinese corpora are relatively reliant on word orders, whereas the Japanese-trained model exploits morphological cues to infer the syntactic categories.**

***Index Terms*—Computational model, cross-linguistic difference, hidden Markov model (HMM), learning of word meanings, syntactic category.**

I. Introduction

A CRUCIAL aspect in understanding the mechanisms of human linguistic competence is to reveal the developmental process by which children acquire language. At around one year of age, children begin to speak words, and by the age of two, they typically produce multiword sentences. Although their utterances obey several syntactic rules [1], they are unable to apply these rules to a novel word (see [2] and [3]), suggesting that their syntactic categorization is immature. In particular, the verb category is not generalized. Tomasello [4]–[6] hypothesized that the syntactic rules of young children are constructed in an item-based fashion, i.e., they only remember the transitions between specific words. More abstract syntactic categories and their transition rules are gradually learned from language input between two and five years of age [4]–[7].

Imai *et al.* [8] examined whether English-, Japanese-, or Mandarin Chinese-speaking children could infer a target object or action indicated by a novel word using its basic syntactic category (here, either the noun or verb category). Between the ages of three and five years, they observed different developmental changes depending on the children's native language. The children were shown a movie in which a woman performs a novel action using a novel object (standard stimulus). Concurrently, an adult experimenter read a sentence including a novel nonsense word, *dax*. The sentence presented the word in one of three conditions: as a noun, bare verb, or verb with arguments. The sentences in these three were "this is a dax," "daxing," and "she is daxing it," respectively. Two test movies were then presented to the children. In the first, the same woman performed the same action with a new unfamiliar object [action-same (AS) stimulus]. In the second movie, the woman performed an unfamiliar action different from the standard stimulus, but used the same object [object-same (OS) stimulus]. The children were asked to select the movie that represented *dax*. The correct choices were the AS stimulus for the two verb conditions and the OS stimulus for the noun condition. The results demonstrated that regardless of the native language, three-year-old children could successfully choose the OS stimulus for the noun condition, but were unable to generalize the verb to the AS video for the two verb conditions [8]. Five-year-old children tended to correctly generalize the novel verb to the AS video, but the condition in which the children succeeded in generalizing the verb was different across the three language groups. The Japanese-speaking five-year-old children were able to correctly generalize the verb to the AS video in the bare verb condition. In contrast, the correct choice percentages for the English-speaking children were consistent with levels of pure chance, and the Chinese-speaking children generalized the verb to the OS video in

Manuscript received June 26, 2017; revised March 9, 2018 and September 13, 2018; accepted November 14, 2018. Date of publication November 23, 2018; date of current version September 9, 2020. This work was supported in part by JSPS Grants-in-Aid for Specially Promoted Research under Grant 24000012, in part by Grants-in-Aid for JSPS Fellows under Grant 13J00756, and in part by JST CREST, Japan under Grant JPMJCR17A4. *(Corresponding author: Yuji Kawai.)*

Y. Kawai and M. Asada are with the Graduate School of Engineering, Osaka University, Suita 565-0871, Japan (e-mail: kawai@ams.eng.osaka-u.ac.jp; asada@ams.eng.osaka-u.ac.jp).

Y. Oshima was with Osaka University, Suita 565-0871, Japan. He is now with NTT Software Innovation Center, Tokyo, Japan.

Y. Sasamoto was with Osaka University, Suita 565-0871, Japan. He is now with Fujitsu Finland Oy, Helsinki, Finland.

Y. Nagai is with the National Institute of Information and Communications Technology, Osaka 565-0871, Japan (e-mail: yukie@nict.go.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2883048

the bare verb condition. The English- and Chinese-speaking children only correctly generalized the verb to the AS video for the condition of a verb with arguments. The outcome is that this paper identified developmental changes and cross-linguistic differences in children's verb generalizations (see Appendix A for details of the experimental data we used).

This paper indicates that children's inferences of word meanings rely on syntactic structures in their native languages. To accomplish the task presented by Imai *et al.* [8], [9], children must infer the syntactic category of a novel word, i.e., determine whether it is a noun or a verb. These syntactic categories enable them to select the correct semantic feature in a scene, based on the knowledge that nouns and verbs typically designate object and action categories, respectively. A key mechanism here is the acquisition of syntactic categories. Three-year-old children may exhibit immature syntactic categorization skills, resulting in a lower accuracy in the word generalization task. Five-year-old children are likely to exhibit cross-linguistic differences in their performances, because they have acquired different syntactic categories based on their native language. This raises two open questions. First, which structures of syntactic categories lead to the differences in word generalization? Second, which characteristics of native languages result in the acquisition of these different structures of syntactic categories? Behavioral and experimental approaches have struggled to reveal the mechanisms underlying this process directly from observed data.

In order to address both of these questions concurrently, we propose a computational model that learns word meanings based on syntactic categories acquired from an input language. The model can examine the acquired representation of the syntactic categories and the word meanings learned through the categories. We evaluate our model by comparing its output with children's word generalizations reported by Imai *et al.* [8]. We employ a hidden Markov model (HMM), which is an unsupervised machine learning technique, to learn syntactic categories [10]. Given certain word sequences, the HMM determines the syntactic categories (hidden states) of words and the transition probabilities between categories. A key mechanism for explaining performance differences between age groups is the number of syntactic categories, i.e., hidden states of the HMM. The larger the number of hidden states the model has, the more complex syntactic rules it can acquire, thus achieving more accurate inferences of word meanings (the first issue). This factor may represent the consequence of child development. We test the performance of the model in three language environments (English, Japanese, and Chinese), and identify how the differences among syntactic structures in these languages affect the acquisition of syntactic categories, and how such language-specific syntactic categories impact the performance of inferring word meanings (the second issue).

We expect the model to behave differently depending on the language, in the following aspects.

1) In English, owing to stronger syntactic cues (i.e., word orders) in categorizing verbs, our model mainly learns word orders. In particular, a verb is positioned after a noun. Hence, it can infer a novel verb as an action if its argument structure is followed (i.e., in the condition of a verb with arguments). Otherwise, it cannot make an inference, even when the model has a sufficiently large number of categories.
2) In Japanese, the trained model can correctly infer an action even for the condition of a verb without arguments, because Japanese contains important morphological cues for verbs, e.g., a verb stem and suffix "teiru" for the progressive form. Japanese contains comparatively weak syntactic cues, and arguments are often dropped.
3) In Chinese, there are no cues for assigning categories to bare words. Therefore, the Chinese-learned model cannot infer the meanings of verbs for the condition of a bare verb, even with a large number of categories.

We will demonstrate that our model behaves as described above. In addition, the acquired syntactic categories are analyzed in order to investigate the origins of language-specific developmental changes in inferring word meanings.

II. Related Studies

A. Cues for Syntactic Categorization

Many models and analytical studies have addressed children's acquisition of syntactic categories as explained below. These suggest that children employ three main cues to induce syntactic categories, i.e., syntactic (distributional), morphological, and phonological.

Almost of all computational models consider syntactic or distributional information, i.e., patterns of ordered pairs of words (e.g., [11]–[19]). For example, given sentences such as "they use balls" and "they throw balls," we can see that the words adjacent to "use" and "throw" are the same. Words belonging to a syntactic category tend to share adjacently positioned words, providing a reliable cue for categorizing the words. The accuracy of this categorization is improved by not only the regularity of word orders, but also by that of syntactic category orders, e.g., the category of transitive verbs is placed after the subject category and before the object category (e.g., [12] and [16]). Put simply, the syntactic category of a word can be determined from the regularities of both adjacent words and adjacent categories. Such syntax-based models employ several unsupervised techniques to induce categories. Elman [11] demonstrated that a simple recurrent neural network (SRN) could determine syntactic categories using simple three-word English sentences. The SRN was trained to predict the next word using the current word, resulting in the acquisition of distributional information in the hidden neurons of the SRN. Syntactic categories were obtained by clustering the representation in the hidden neurons. Therefore, the output relied on the previous syntactic category as well as the input (i.e., the previous word). The HMM that we adopt addresses sentences in a similar manner, where the next word is determined using the syntactic categories.

Recent studies analyzing child-directed speech have highlighted the importance of morphological and phonological cues in the acquisition of syntactic categories. Onnis and Christiansen [20] reported that child-directed speech contains rich morphological information that can classify syntactic categories in many languages. For example, English verbs are often accompanied by verb suffixes such as "-ing" and "-ed." An experimental study found that English-speaking 15-month-old children could already segment verbs into bound morphemes and free morphemes [21], [22]. Thus, children seem to employ such morphological cues to infer syntactic categories. Furthermore, it has been suggested that several phonological components may be useful in distinguishing syntactic categories. For example, the mean length of syllables is significantly greater for nouns than for verbs in English [23]. In addition, one modeling approach demonstrated that considering phonological and morphological cues, i.e., the number of syllables and the ending phoneme, can improve the performance of syntactic categorization [24].

In summary, previous computational studies have proposed several algorithms for inducing syntactic categories. These studies suggest that children may categorize words using not only syntactic cues, but also word-internal cues. However, most existing studies have only demonstrated limited aspects of children's language development. To verify the plausibility of computational models, we need to investigate the potential of the models for reproducing diverse aspects of language development, such as the differences between languages.

B. Existing Models for Inferring Word Meanings Using Syntactic Categories

The learning of a word meaning is modeled as the calculation of $P(f|w)$, the probability of some semantic features f appearing in a scene when given a word w. Cross-situational learning calculates mappings by considering many utterance-scene pairs (e.g., [25]–[28]).

The basic idea of inferring a word meaning in the current study is to infer f by considering not only w, but also the syntactic category s of the word, i.e., $P(f|w, s)$. Even if w is a newly learned word, s helps the learner to map the word to an appropriate semantic feature, such as in a situation that corresponds to the experiments of Imai *et al.* [8], [9]. Alishahi and Fazly [29] demonstrated that a model for inferring word meanings via syntactic categories improves the accuracy and speed of learning word meanings compared to a model without s. However, they assumed that s was completely correct and given in advance. Therefore, their model cannot explain how s is acquired, or how s influences developmental changes depending on the input language in $P(f|w, s)$. Other models acquire s using corpus-based unsupervised machine learning techniques [19], [30]–[33]. In order to obtain s, Toyomura and Omori [30] employed an SRN where the syntactic representation in the hidden neurons is associated with observed semantic features. Their model can infer the target object indicated by the word through the syntactic representation. Alishahi and Chrupała [32] employed a statistical unsupervised model to learn s, which may be a more reasonable method for modeling complex syntactic structures. They demonstrated that the learning efficiency of the model is similar to that of the model using the supervised syntactic categorization [29]. Attamimi *et al.* [33] showed that syntactic categorization using an HMM can help associate between words and semantic features. The hidden states represent a category of functional words and semantic properties of words such as an object, motion, and place. These information enables the model to accurately infer word meanings. However, they did not consider cross-linguistic differences or developmental changes in terms of the categorization of nouns and verbs. Most of the above-mentioned models have only been evaluated in terms of their learning efficiency, rather than with respect to language dependencies or developmental changes. A similar problem is observed in studies regarding syntactic categorization. The efficiency of a system does not always guarantee its plausibility as a cognitive mechanism.

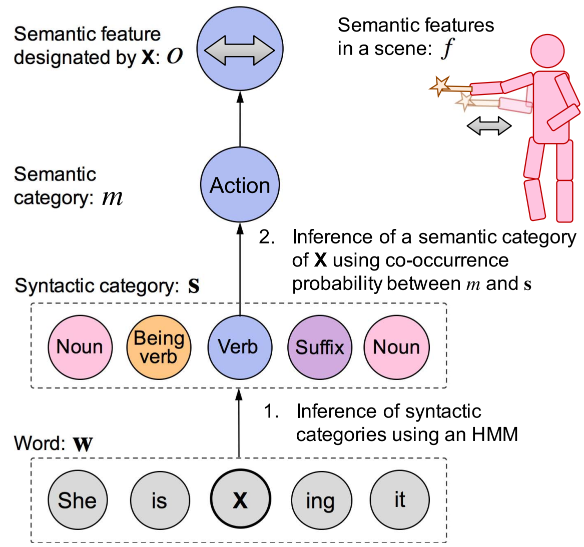


Fig. 1. Overview of the model for inferring word meanings via syntactic categories. This indicates the inference process of a target semantic feature o designated by a novel word X (w_t) through the syntactic category $\mathbf{s}$ and semantic category m. There are two inference processes. At first, the model infers $\mathbf{s}$ of a given word sequence $\mathbf{w}$ using an HMM. Then, the model infers m of X using co-occurrence probability between m and $\mathbf{s}$. The model thus chooses o from observing semantic features f based on m, where it is assumed that the relationship between o and m is given.

III. Word-Learning Model Based on Syntactic Categories

Fig. 1 presents an overview of our model for inferring the probability of a visual semantic feature $P(o)$ designated by a word w_t, for a given word sequence $\mathbf{w} = (w_1, w_2, \ldots, w_t, \ldots)$ and observing semantic features f. It is assumed that $\mathbf{w}$ is a word sequence containing a verb suffix separated from a verb stem. Furthermore, f denotes a set of semantic features corresponding to nouns, verbs, and adjectives. The purpose of the model is to compute $P(o|w_t, f)$. The model selects a target semantic feature o, which is designated by a given word w_t, from the feature candidates f in a given scene. This probability is proportional to $P(o|w_t)P(o|f)$, where it is assumed that w_t and f are independent. Here, $P(o|f)$ is given as a uniform

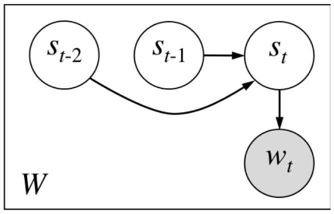


Fig. 2. Generative model of a trigram HMM [34]. A word is generated by a syntactic category that is generated by the previous two categories. W indicates the number of words in a corpus, i.e., the length of a corpus.

distribution over the semantic features present in the currently observed scene. This assumption means that a target feature exists in the observed scene and the model does not have semantic bias to infer the target. The model learns $P(o|w_t)$ using many pairs of scenes and words. However, the model cannot infer a semantic feature o designated by a newly given word w^{novel}, because it has not yet acquired the direct relation $P(o|w_t = w^{\text{novel}})$. Therefore, the model indirectly infers o in two stages.

1) Inference of the syntactic categories $\mathbf{s}$ of $\mathbf{w}$ using an HMM.
2) Inference of the semantic category m of s_t using co-occurrence probability between semantic and syntactic categories.

Finally, the model chooses o from f based on m. Here, $\mathbf{s} = (s_1, s_2, \dots, s_t, \dots)$ denotes the sequential syntactic categories of $\mathbf{w}$, and $\mathbf{s}_{-t}$ denotes $\mathbf{s}$ excluding s_t.

In the first stage of inference, we employ an HMM to derive the syntactic categories $P(s_t|w_t = w^{\text{novel}}, \mathbf{s}_{-t})$ as its hidden states. The generative model for the HMM is shown in Fig. 2 [34], where W indicates the length of a corpus. We assume a trigram Markov chain in the transitions between syntactic categories. The probability of the current syntactic category is estimated based on the previous two categories, i.e., $P(s_t|w_t, \mathbf{s}_{-t}) \equiv P(s_t|w_t, s_{t-2}, s_{t-1})$. The HMM induces these transition probabilities between the syntactic categories, corresponding to simple syntactic rules. For example, an English-trained model is expected to acquire comparatively high values of $P(s_t = \text{“verb stem”}|s_{t-2} = \text{“noun”}, s_{t-1} = \text{“be-verb”})$.

The model infers o from an estimated category s_t, as $P(o|s_t)$. However, there is no chance of learning the probability for a newly given semantic feature. Note that the children were presented with a novel object and action in the experiments of Imai *et al.* [8]. To address this issue, the model needs the second stage of inference. We introduce a semantic category m, to denote more abstract categories of semantic features, such as action, object, or property. We assume that this category unambiguously corresponds to semantic features, and the correspondence is given *a priori*. The distribution of $P(o|m = m_i)$ is uniform over semantic features that belong to the category m_i. The model cannot learn the direct mappings between $\mathbf{s}$ and the novel semantic feature o in advance. Nevertheless, it can learn the mappings between $\mathbf{s}$ and m. We assume that the model can recognize m for the novel o. Thus, the newly given o can be indirectly inferred from $\mathbf{s}$ through m, as $P(o|s_t) = \sum_m P(o|m)P(m|s_t)$.

The probability that o is referred to by w_t is obtained as the product of the direct pass $P(o|w_t)$ and the indirect one $\sum_{s_t} \sum_m P(o|m)P(m|s_t)P(s_t|w_t, \mathbf{s}_{-t})$

$$P(o|w_t, \mathbf{s}_{-t}, f) \propto \sum_{s_t} \sum_{m} P(o|m)P(m|s_t)P(s_t|w_t, \mathbf{s}_{-t})P(o|w_t)P(o|f) \quad (1)$$

where $P(o|w_t)$ on the right-hand side denotes a direct mapping from a word to semantic features. This probability is uniform if w_t is novel. Furthermore, we divide the right-hand side of (1) by $\max_o(P(o|w_t))$ to ensure that a novel word readily corresponds to a novel semantic feature. This assumption is an example of the mutual exclusivity that a newly given word should correspond to a novel visual feature [35].

The model must learn two probabilities in (1). First, $P(s_t|w_t, \mathbf{s}_{-t})$, i.e., the induction of the syntactic category of a novel word on the basis of the transition rules between the categories, and second $P(m|s_t)$, i.e., the correspondence between the syntactic and semantic categories. As mentioned above, we employ an HMM to determine $P(s_t|w_t, \mathbf{s}_{-t})$. In this HMM, learning is based only on a learning corpus $\mathbf{w}$, and does not rely on o or m (see Appendix B.A for further details). In addition, $P(m|s_t)$ is obtained in a cross-situational manner. Given many pairings of a sentence with various semantic features, the probability is calculated based on co-occurrence between the syntactic and semantic categories in each sentence-feature pair (see Appendix B.B for further details).

A key mechanism of the model is the inference of syntactic categories, $P(s_t|\mathbf{w}, \mathbf{s}_{-t})$. This probability is acquired based on a language corpus (see the next section for details of the corpora employed for the learning). Thus, the syntactic categories and their transition rules rely on the syntactic structures in the given language. The number of hidden states S in the HMM is a given constant for each age simulation. We suggest that the number has a crucial influence on the accuracy of the estimated syntactic category. If S is small, such as two, then the hidden states may only represent a start-word category and a category for other words, as illustrated in Fig. 3(a). Such immature categories lead to inaccurate estimations of the probabilities of o and m. In contrast, a model with sufficiently large S can correctly infer m and o, because the verb category is suitably differentiated from the noun category [see Fig. 3(b)]. We hypothesize that different Ss can replicate the manners in which children learning novel nouns and verbs at different ages, as reported by Imai *et al.* [8]. We expect models with smaller and larger values of S to reproduce the behaviors of three- and fiver-year-old children, respectively.

IV. Experimental Setting

An experiment was conducted to investigate whether the model can replicate children's generalizations of nouns and verbs [8], and to clarify the origins of the observed differences between ages and three languages. The model was trained using artificial utterance-scene pairs, and then evaluated by inferring a target semantic feature designated by a novel word. Subsequently, the internal representation, i.e., the hidden states of the model, were analyzed to reveal the syntactic categories

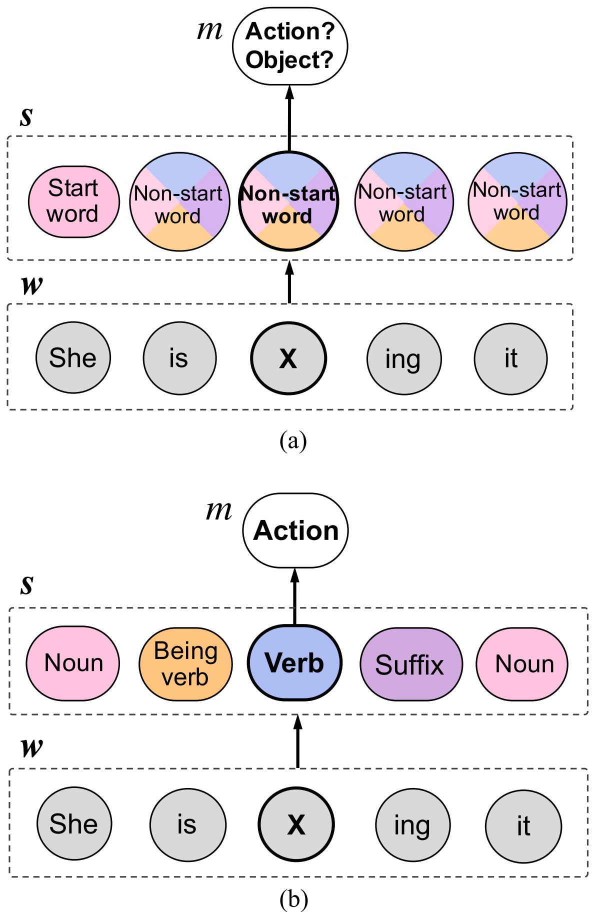


Fig. 3. Inference of word semantics depending on the number of syntactic categories. (a) Model acquires obscure syntactic categories because the number of syntactic categories S is smaller, resulting in the failure to infer a target semantic feature designated by a novel word. (b) Sufficiently large S enables the model to acquire well-differentiated syntactic categories. The larger S improves the accuracy with which target semantic features are inferred. (a) Estimation through nondifferentiated syntactic categories. (b) Estimation through well-differentiated syntactic categories.

that contributed to the inference of meanings of novel nouns and verbs. The following sections describe the learning and evaluation processes of the model.

A. Utterance-Feature Pairs for Learning

We created artificial corpora containing the syntactic and morphological structures of English, Japanese, and Mandarin Chinese, to be used as learning datasets. To avoid any semantic influences, we employed these artificial corpora instead of real ones to focus on the acquisition of syntactic categories. Probabilistic transition rules between word categories (e.g., subject, object, verb, object, and adjective categories) were designed for each language, and words were selected from the vocabulary of each category in a random manner. The word categories and their transition probabilities were determined based on real corpora, consisting of utterances made by caregivers to children aged between two and five years [36]–[38]. These were obtained from the CHILDES database [39]. In particular, the ratios of syntactic structures in the English corpus are identical to those from an analysis of child-directed-speech [40] (see Appendix C for details). For simplicity, no nested sentences are contained in the artificial corpora, i.e., simple sentences such as "the man is reading a book" are given to the model.

In general, subject and object words are dropped from sentences in the Japanese corpus more often than in English and Chinese. Japanese sentences can also exchange the subject with the object, whereas the word orders of the English and Chinese corpora are more stable. The percentage of single sentences consisting of free morphemes of a verb and the progressive tense suffix "-ing," i.e., bare-verb sentences, is very small (0.2%) in the English corpus. In contrast, more bare-verb sentences occur in the Japanese corpus.

An important assumption is that the words in all of the corpora have separate suffixes and free morphemes. For example, progressive verbs were separated into two morphemes (a verb stem and the suffix "-ing") in the English corpus. This coding allows the model to infer syntactic categories based on morphological cues as well as syntactic cues. This assumption is plausible, because young children already have the ability to segment a verb into its bound morpheme and free morpheme [21], [22].

The scenes corresponding to utterances in each corpus were fed to the model simultaneously. These scenes artificially consist of semantic features designated by nouns, verbs, and adjectives in the utterances. In addition, a property is virtually assigned to a noun unless that noun is qualified by any adjectives in the text. Hence, we assumed that all objects and agents have properties of some types. For example, for the utterance "the man is reading a red book," the semantic features "property of the man," "man," "read," "red," and "book" are observed by the model. We adopted three types of semantic categories: object, action, and property. They unambiguously correspond to nouns, verbs, and adjectives, respectively. It was assumed that the model knew the allocation of semantic categories to semantic features. (Categorization of semantic features is also a big issue in language development, but it is not addressed in the present study.) In this experiment, the model learned 10 000 utterance-scene pairs.

B. Evaluation

A sentence containing a novel word X that did not occur in the learning corpora was provided to the model following the learning process. We examined whether the model could estimate the correct target semantic feature for X under three conditions regarding sentences containing X. This constitutes the same experiment as that of Imai *et al.* [8], [9] (see Appendix A for details).

1) English
   a) Noun condition: This is an X.
   b) Bare verb condition: X-ing.
   c) Verb with arguments condition: She is X-ing it.
2) Japanese
   a) Noun condition: X ga (nominal particle) aru (exist).
   b) Bare verb condition: X teiru (verb suffix: progressive).

c) Verb with arguments condition: Oneesan (noun: girl) ga (nominal particle) nanika (noun: something) wo (accusative particle) X teiru (verb-suffix: progressive).

3) Mandarin Chinese

a) Noun condition: Nali (noun: there) you (verb: exist) ge (classifier) X.

b) Bare word condition: X.

c) Verb with arguments condition: Ayi (noun: girl) zai (adverb: progressive) X yi (one) ge (classifier) dongxi (noun: thing) ne (mode marking particle).

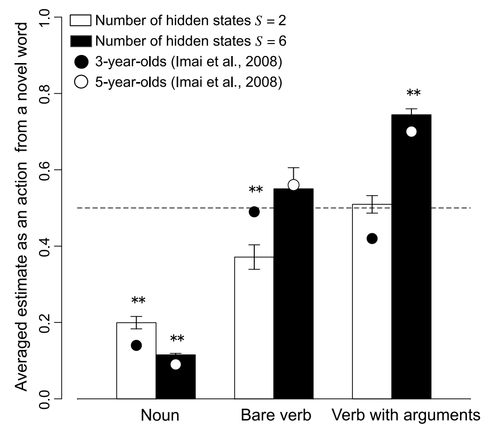
The model also observes five semantic features that denote "property of the girl," "girl," "a novel object," "a novel action," and "property of the novel object" in all conditions. The properties of the girl and the novel object were randomly assigned to adjectives of the corpora's vocabulary. The model then estimated the probabilities that X refers to semantic features o in the utterance, based on (1). The estimated probabilities for the semantic features other than the novel object and action, i.e., property of the girl, girl, and property of the novel object, were equally given to the probabilities for the novel object and action. This additional operation is consistent with the children's forced choices between the OS and AS stimuli, even if they selected semantic features other than the novel object and action as the target.

We evaluated 20 variations of the model with different initial hidden states, and averaged the conditional probabilities for the novel action referred to by X under each condition. A t-test (two-tailed with $\alpha = 0.05$) was employed to evaluate the accuracy of the estimation. Values that are significantly smaller than a chance level (0.5) are considered correct in the noun condition, and significantly larger values are correct in the verb conditions. We compared these values for in the cases of three- and five-year-olds, which have smaller and larger numbers of hidden states S, respectively. The numbers were determined in order to fit the model performance to that of the children in the experiment of Imai *et al.* [8], resulting in two states for the three-year-old case and between four and six states for the five-year-old case (see Appendix D for details).

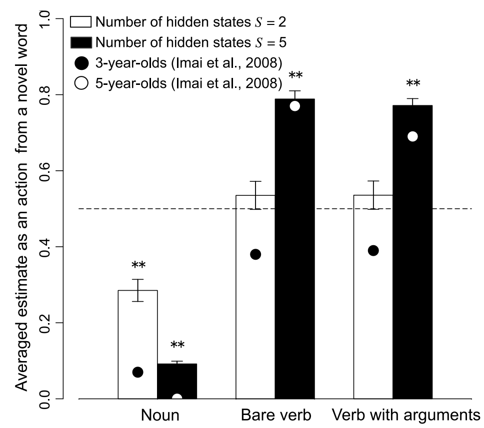
Furthermore, we analyzed the acquired hidden states to investigate the structures of the syntactic categories. We tagged the learned corpora with their parts of speech, and investigated which parts of speech were represented by the hidden states. For example, the value for the representation of verbs is given by

$$\sum_{i \in \text{“verbs”}} P(s_i | \mathbf{w}, \mathbf{s}_{-i}) \qquad (2)$$

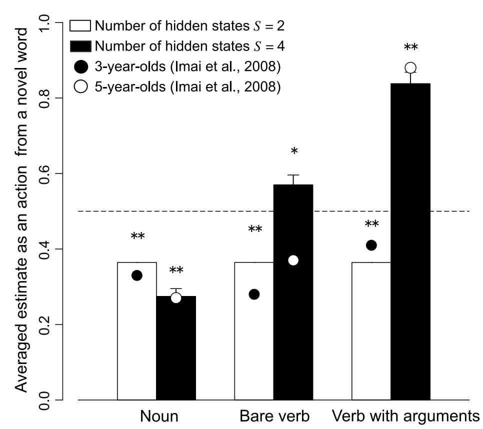
where "verbs" is a set of indexes of verbs in the corpus. The percentage of values for all parts of speech was then calculated in each hidden state.



(a)



(b)



(c)

Fig. 4. Estimation of the target semantic feature corresponding to a novel word by the model (bars) and children (points). Results for the model trained with the (a) English, (b) Japanese, and (c) Chinese corpus. The bars indicate the average percentage of model estimates in which the inferred semantic feature was a novel action. The white and black bars denote the results for the models with fewer hidden states (three-year-old case) and more hidden states (five-year-old case), respectively. The stars above the bars denote significant differences (*: $p < 0.05$, **: $p < 0.01$) between the model estimation and the chance level (0.5), shown as a dashed line. The error bars indicate the standard error. The black and white points denote the results achieved by three- and five-year-old children, respectively, as reported by Imai *et al.* [8].

V. Results

A. Inferring the Meaning of Novel Word

Fig. 4 presents the experimental results for the (a) English, (b) Japanese, and (c) Chinese language conditions. The vertical axis indicates the percentage with which the novel action was selected as a target semantic feature of the novel word. Therefore, the value in significantly more than the chance level (the dashed line) exhibits that the target is corresponded to the action. Conversely, the significantly lower value than the chance level indicates that the target is corresponded to the object. The bars and circles denote the performance of the model and the experimental results reported by Imai *et al.* [8], respectively. The white and black bars indicate the results for the three- and five-year-old cases, i.e., those with smaller and larger values of S, respectively. The solid and empty circles indicate the results from children aged three and five years, respectively.

For all of the conditions, estimations of the model were similar to the children's results, suggesting that the model is able to replicate the language-specific differences in inferring the meanings of novel nouns and verbs in terms of age. Both the model and the children correctly selected the novel object in the noun condition, regardless of their language and age (all with $p < 0.01$). The values in the three-year-old cases (with $S = 2$) were less than or close to a chance level for both verb conditions. In contrast, the results for the five-year-old cases exhibit differences depending on the learned language for the verb conditions. Although the English-trained model with six hidden states successfully selected the novel action for the condition of a verb with arguments ($p < 0.05$), it did not succeed in the bare verb condition ($p > 0.05$) [see Fig. 4(a)]. The averaged results for the Chinese-trained model (with $S = 4$) indicate a similar tendency to the English-trained model, but the estimation value for the novel bare verb was significantly higher than the chance level ($p < 0.05$) [see Fig. 4(c)]. In contrast, in the case of five-year-olds for Japanese with five hidden states ($S = 5$), the novel verbs could be properly mapped to the target action for both of the verb conditions ($p < 0.01$) [see Fig. 4(b)]. However, the percentage of the actual children's choices for the condition of the verb with arguments was not statistically significant compared to the chance level [8], although the averages of the model's and children's estimations exhibited similar trends.

B. Representation of Syntactic Categories

To understand how the differences arising between the ages and languages in the above results arise, we analyzed the part-of-speech representation of the hidden states of the HMM using (2). Fig. 5 illustrates a typical representation of the (a) English, (b) Japanese, and (c) Chinese cases. The graphs on the left and the right represent the three- and five-year-old cases, respectively. The colored blocks denote the ratios of the part-of-speech representations. Parts of speech that are common between the languages include nouns (red), transitive verbs (deep blue), intransitive verbs (blue), and adjectives (orange). Other parts of speech are defined according to the syntactic structures (see Appendix C for a detailed explanation of parts of speech).

In most of the English-trained three-year-old cases, the model acquired hidden states that separately represent nouns (ID 1) and verbs [ID 2 on the left of Fig. 5(a)]. The noun category enabled the model to predict the semantic feature of the novel noun correctly. However, the verb category included the adjectives and part of the nouns, resulting in chance-level estimations in the verb conditions. In contrast, the model acquired a well-differentiated part-of-speech representation for the five-year-old case [on the right of Fig. 5(a)]. The model was able to infer the target action for the novel verb with arguments using the verb category (ID 3 and ID 4). The noun category was separated into an object-words category (ID1) and a start-words category, i.e., a subject category (ID 2), suggesting that the structure of the syntactic categories depends on the word order. Thus, given the bare verb, the model regarded the start word as a noun, although the following word was the progressive suffix. Such competition between syntactic and morphological cues resulted in chance-level estimates in the bare verb condition.

There were two representations in the Japanese-trained three-year-old case. First, the category of open-class words (nouns, verbs, and adjectives) (ID 1) and the category of closed-class words (ID 2) were acquired, as shown on the left of Fig. 5(b). These undifferentiated categories resulted in the estimation being almost at a chance level in all conditions. This occurred 11 out of 20 times. In the other representation, the nouns and verbs were represented separately. As the verb category also included adjectives, the average performance of the model was at a chance level. In contrast, one category in the five-year-old case only represented verbs [ID 2 on the right of Fig. 5(b)]. The novel verbs were classified into the verb category, because the model acquired the rule that this category is arranged before the category of verb suffixes (ID 5). Japanese sentences often drop subject and object words and change the word order, resulting in the acquisition of syntactic categories that rely on word suffixes, i.e., morphological cues rather than syntactic ones. Therefore, competition between cues, as was present in the English-trained model, did not occur in the Japanese-trained model, which was able to infer the correct target semantic feature in the bare verb condition.

In the Chinese-trained three-year-old case, category ID 1 [on the left of Fig. 5(c)] consisted of start words, such as nouns, adjectives, and verbs, and words following words belonging to the other category (ID 2). The other category included adverbs and verbs without any adverbs. Hence, the novel verbs were allocated to category ID 1, because they were located at the beginning of a sentence or after an adverb, resulting in estimation at close to a chance level. In the five-year-old case, the model acquired a category consisting of only verbs [ID 2 on the right of Fig. 5(c)]. However, the Chinese-trained model was unable to classify the bare word into the verb category, because it did not acquire any categories representing verb suffixes, unlike the English- and Japanese-trained models, i.e., there were no word-internal cues. When the novel verb was placed after a word from the adverb category (ID 4), the model could correctly map the novel word to the action feature.

C. Development of Syntactic Categories: Predictions From the Model's Approach

Fig. 6 illustrates the contents of the hidden states in terms of S. We set the concrete number of hidden states in each trial because our model cannot gradually increase this number during the learning. The corpus was fed into the model to learn it

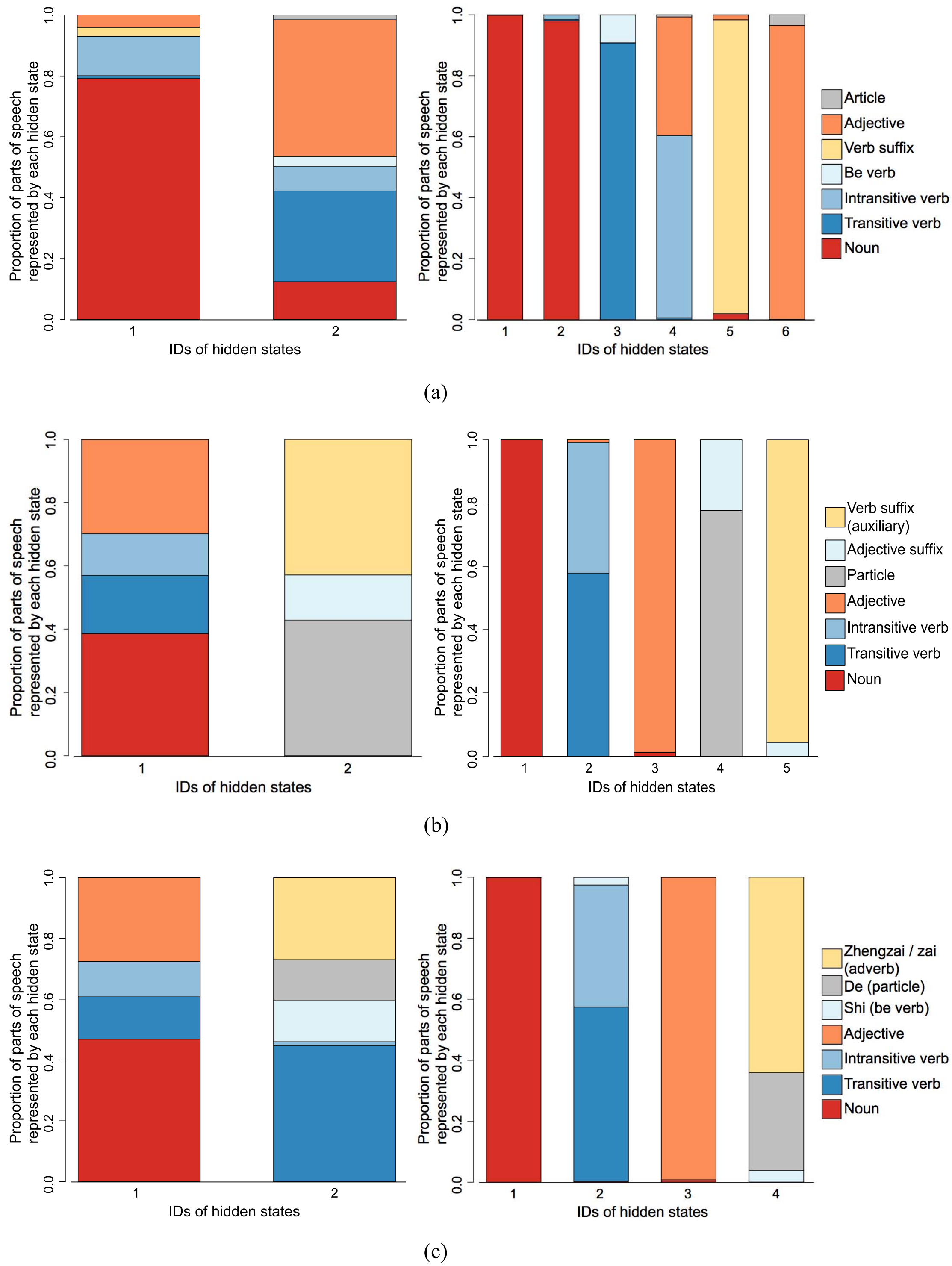


Fig. 5. Part-of-speech representation for each hidden state of the HMMs. These HMMs were learned the (a) English, (b) Japanese, and (c) Chinese corpus. The graphs on the left indicate a typical representation of the three-year-old cases, whereas the graphs on the right indicate a typical representation of the five-year-old cases.

in the same manner as previous sections. We hypothesize that changes in these contents may correspond to the development of syntactic categories in children.

The result shows that the model roughly represented main grammatical structures, e.g., relation between a subject and object, in the smaller number of S. As the number became larger, the model could capture more fine-grained grammatical structure, e.g., a suffix and particles. This complements the results at the ages of three and five in the previous sections. In the English-trained model [Fig. 6(a)], a noun category was generated first. This category was then differentiated into a category of subject nouns (initial words) and

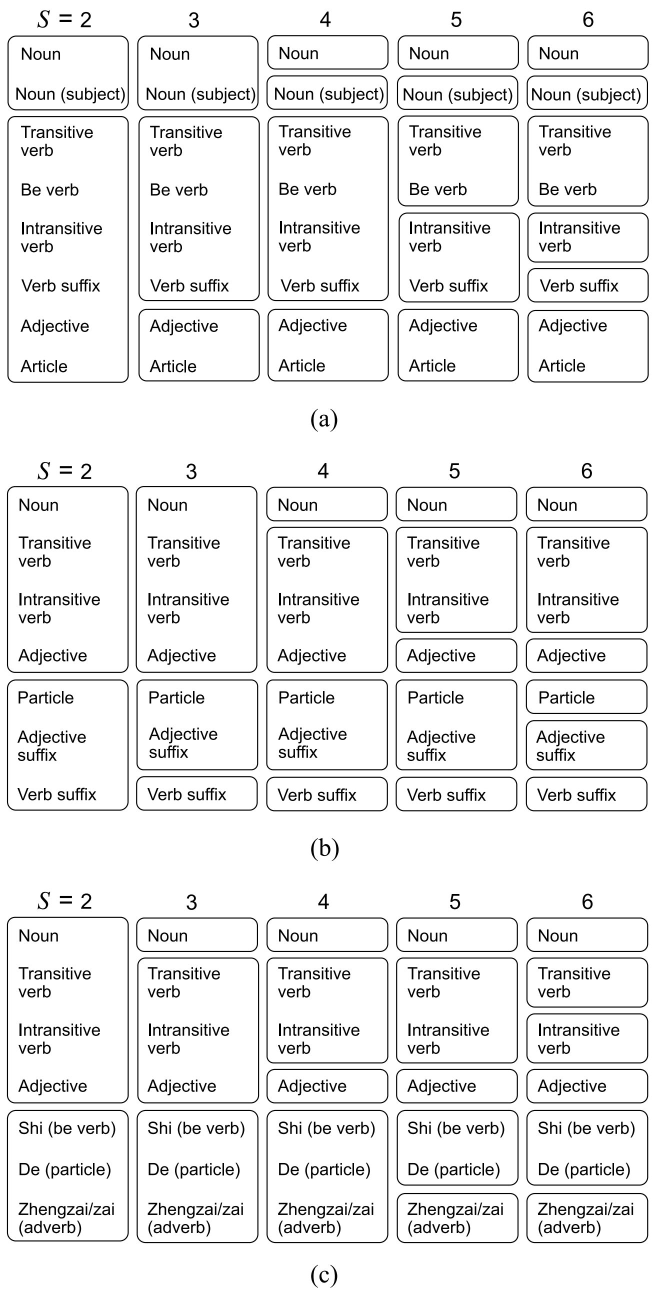


Fig. 6. Differentiation processes of syntactic categories. Different part-of-speech representation of the hidden states with the number of hidden states varying from two to six in the (a) English, (b) Japanese, and (c) Chinese cases.

one of noninitial words. This result suggests that the process of acquiring English syntactic categories relies on the word order, i.e., syntactic cues. In contrast, the Japanese model displayed differences between closed- and open-class word categories [see Fig. 6 (b)]. The closed-class word category separated a suffix category from a particle one, suggesting that this acquisition process is based on the particles and suffixes. Similarly, the Chinese-trained model [Fig. 6 (c)] produced a category of closed-class words. However, this model did not represent suffixes, and the category of closed-class words was subdivided into adverb and particle categories. The adverbs are important syntactic markers, because a word can be both a noun and a verb in Chinese.

VI. Discussion

We have proposed a model that may represent children's inferences of novel nouns and verbs. Namely, an HMM derives the syntactic categories of words from the input sentence, and these categories are associated with observed semantic features. The proposed model then infers an appropriate semantic feature that is referred to by a novel word using the acquired syntactic categories. The model was evaluated by comparing its output with the performance of children in English, Japanese, and Chinese environments, as reported by Imai *et al.* [8]. The experimental results demonstrate that the model can successfully replicate the language-specific differences in the performances of the children, in terms of age (see Fig. 4). Further analysis of the hidden states in the learned HMMs revealed a detailed representation of the syntactic categories that yielded the differences in performance in the three languages (see Fig. 5). Our findings prompt the following three major suggestions.

A. HMM as Model Toward Syntactic Development

First, the proposed model offers a computational framework for explaining young children's syntactic categorization from language input. We have also demonstrated that these abstract syntactic categories may affect the learning of words by children. There is strong evidence that a single model could replicate the multiple tendencies in children's inferences of word meanings in the three language environments for the two age groups considered here. Generalizing syntactic knowledge to a novel word requires abstract syntactic categories and schemas, which seem to develop from around two to three years of age [4]–[6]. Tomasello also emphasized the importance of language input for the construction of a child's syntactic knowledge. Connectionist studies agree with this view, i.e., language acquisition is regarded as the learning of several statistical and probabilistic properties in the language children are exposed to [41] and [42].

In the present model, hidden Markov states constitute a key structure for representing syntactic categories. Note that structures of this type have been widely applied to computational models of cognitive competence and artificial intelligence, such as in action learning [43] and the theory of mind [44]. This suggests that the framework of our model would not be specific to the language competence, which contrasts with the strict universalist view that only accounts for the human language as universal grammar. However, our model is too simple to explain some linguistic properties, such as syntactic hierarchies. From the viewpoint of natural language processing, the unsupervised learning of this hierarchical structure is based on part-of-speech categorization (see [45]). Therefore, our model might be considered as a primitive syntactic structure.

Our model simply investigated the effects of linguistic (morphosyntactic) properties on word learning. However, we do not claim that linguistic factors exclusively determine the learning of word meanings in the early stages of lexical development.

The acquisition of a lexicon is supported by multiple factors, i.e., linguistic and extralinguistic (perceptual and semantic) factors that describe categories of words and semantic features. In fact, Imai *et al.* [8] highlighted the importance of extralinguistic cues in learning the meanings of novel words (see also Appendix A). In particular, the Chinese-speaking five-year-old children begin to incorrectly select a target object for the verb condition if the object was more salient in the standard stimulus movie. An open question is how children integrate linguistic and extralinguistic cues, i.e., how the model acquires the semantic categories from visual input and processes interactions between the syntactic and semantic categorizations to construct a lexicon. Our model can easily adapt with such semantic constraints because of its simplicity. A recent robotic study proposed a system for acquiring word meanings by categorizing raw visual and auditory information sensed by a robot in a real environment [33], [46]. By applying such a technique, we plan to integrate the perceptual and syntactic mechanisms, in order to propose a more plausible model for lexical development.

B. Developmental Change

The differences in the inferences of word meanings for different ages can be explained by the different numbers of hidden states of the HMM. When this number was small, the model acquired ambiguous syntactic categories in the hidden states. The acquired noun category was relatively unambiguous, but the verb category included words belonging to other parts of speech in all of the three languages. This is because nouns have more distributional cues for finding the correct syntactic category universally, e.g., a word at the beginning of a sentence. Therefore, the model was able to infer that an object feature was referred to by a novel noun although it did not correctly choose an action for a verb. In contrast, a larger number of hidden states enabled the model to obtain well-differentiated syntactic categories, resulting in successful inferences of novel verbs. We employed different numbers of syntactic categories for the five-year-old cases for different languages, i.e., $S = 6$ for English, $S = 5$ for Japanese, and $S = 4$ for Chinese. These numbers were chosen to fit the resulting performances to those presented by Imai *et al.* [8]. The differences may reflect different levels of grammatical complexity in the learning corpora. Therefore, the absolute values of these numbers are not of key importance, but rather we emphasize that the different numbers can explain the inferences of word meanings by children at different ages.

Constructing a mechanism to increase this number, i.e., a trigger for syntactic development, remains an open issue, because the current model requires the number to be set in advance. We could investigate advantages of starting small [47], [48] which if the model automatically and gradually developed the categories. In accordance with this hypothesize, starting learning of word meaning from a small number of categories might facilitate the learning. However, the current model could learn syntactic categories well even if it starts to learn the categories from the large number of categories.

There are two potential types of trigger: 1) input-independent and 2) input-dependent factors. The first includes brain growth, which is mostly independent of linguistic input. A child's brain develops rapidly between the ages of three and five years (e.g., [49]–[51]). Although some brain networks are organized by sensory information, genetic mechanisms appear to be a strong factor in brain maturation. The developmental mechanism in the current model assumes that this automatic brain growth is independent of any linguistic and sensory input.

Regarding an input-dependent factor, many studies have indicated that verb-argument structures in utterances by young children are similar to those of their caregivers (see [52] and [53]). Therefore, it is widely accepted that a caregiver's language input is an important factor in the development of a child's syntactic competence. Speech directed to younger children has been found to be simpler than that directed to older children (see [54] and [55]). Snow [54] compared utterances to English-speaking two-year-old children with those to English-speaking ten-year-old children. The mean length of utterances to the two-year-old children was found to be shorter, and the utterances consisted of fewer compound verbs and subordinate clauses. It was asserted that such grammatically simplified speech is tractable for the younger children, which aids their language learning [54]. We hypothesize that exposure to grammar of an increasing complexity as a child becomes older may be a trigger for syntactic development. A small number of hidden states in the HMM is sufficient for representing simple constructions. Additional hidden states are required to model the complex sentences that are conveyed to older children. It is known that the model parameters, including the number of hidden states, can be optimized to provide a better representation of the language input. An HMM-based machine learning method to optimize the number of hidden states has been proposed [56]–[58]. In the future, we plan to employ such a model to verify our hypothesis that the complexity of language input is one of the triggers of syntactic development.

To begin with, how a child forms semantic categories is also a big problem although we assumed that they were given in advance. The well-known hypothesis about semantic categorization is the noun predominance, i.e., the learning of nouns is easier and faster than that of verbs (see [59]–[62]). Some observational and experimental studies have assigned the origin of noun predominance to perceptual and semantic factors. Namely, it is more difficult to parse actions than objects, because the spatiotemporal boundaries of actions are not clear [8], [59], [60], [63]. Such semantic categorization may affect syntactic development. Further research is needed to understand interactively development of semantics and syntax.

We made several assumptions on the given utterances and scenes: The utterance had only simple but typical grammatical structures (see Appendix C) and, similarly, the scenes consisted of the limited features corresponding to the utterances. These assumption enhanced the learnability of the model and be considered as scaffolding from caregivers [64]–[66]. Furthermore, not only the scaffolding but also social interactions between caregivers and learners are definitively important for language acquisition [67]. For example, the caregivers should change ways of emphasizing a target object and speech depending on the learners' behavior, e.g., attention. It is expected that we will extend the model so as to adapt such scaffolding to the learner's behavior and explore how to design the interactions.

C. Cross-Linguistic Difference

Cross-linguistic differences in generalizations for verbs in the five-year-old case were the result of different processes to acquire categories between the input languages. English contains more morphosyntactic cues for inferring syntactic categories, i.e., syntactic rules for the English corpus are more stable. That is, change in word orders and dropped subject or object words rarely occur. In addition, English verbs are often with bound morphemes. Therefore, if the one-word sentence "X-ing" was given, the model inferred the syntactic categories of X using two cues: 1) an initial word or a word following an article is a noun and 2) a word preceding "-ing" is a verb. Consequently, for the bare verb condition the result was at a chance level, because of conflicts between the cues.

However, syntactic cues, i.e., subject and object words, are often omitted in the Japanese corpus although Japanese verbs have similar morphological cues as English verbs do. The syntactic categories learned from the Japanese corpus relied on morphological cues, resulting in the correct inference of a verb category in the bare verb condition. On the other hand, there are no verb suffixes in the Chinese corpus, therefore, word order (e.g., an adverb coming before a verb) is a relatively strong cue for identifying the verb category. Thus, the Chinese-trained model could not successfully determine the verb category when a novel bare word was provided without any syntactic cues. Consequently, the model used proper cues to acquire the syntactic categories. If the test sentences included cues, then the model could identify the correct category for a novel word, resulting in a successful inference of its meaning. Although previous studies have pointed out the computability of syntactic categories using several cues, as commonly observed in many languages (see [17], [20], [23]), our model suggests that children may use suitable cues for their native language, as demonstrated in the experimental study [8]. This mechanism reminds us of the competition model (see [68], [69]), positing that a word meaning is inferred through a competition between many cues within a sentence. A competitive mechanism is not explicitly embedded in our model. Nevertheless, our model automatically selects more advantageous cues in identifying syntactic categories through learning. Our approach provides new insight into how the competitive mechanism develops in children.

VII. Conclusion

We have proposed a computational model that can reproduce the inferences of meanings of novel words by children at different ages with different languages, as reported by Imai *et al.* [8]. Employing different numbers of hidden states made it possible to represent children's performances at different ages. The model acquired language-specific syntactic categories, which produced the observed cross-linguistic differences. As a result, the estimates provided by the model were similar to the performances of the children. An analysis of the representation of the hidden states revealed the detailed structure of the acquired categories underlying the differences. The English-learned syntactic categories relied on syntactic cues, because English contains stable syntactic rules. The English model consequently failed to select a target action designated by a novel verb without arguments. Japanese sentences often lack syntactic cues, such as subject and object words, resulting in a stronger influence of morphological cues for the acquisition of Japanese syntactic categories. Therefore, the Japanese model was able to map a novel bare verb to an action correctly. In contrast, Chinese does not contain any morphological cues, but does have syntactic cues. Thus, the Chinese-trained model was unable to determine the reference of a bare word in a significant proportion of cases. In addition, by using our model we deduced a hypothesis regarding the processes underlying the development of children's syntactic categories based on the number of hidden states, which corresponds to the number of syntactic categories.

Our model can generate testable predictions for language development study. It suggests that children select more advantageous cues (e.g., word order and morphology) to infer word meanings as the result of grammatical statistical learning. Children's behavior given the competitive cues are interesting for a benchmark of cognitive models as well as research of developmental psycholinguistics. Further computational and experimental investigation is expected to verify this hypothesis.

Appendix A
Data Cited From Imai *et al.* [8]

We used experimental data provided by Imai *et al.* [8] to compare with our simulation results. The Japanese, English, and Chinese data sets were borrowed from the results of Study 1 (Table 3), Study 2 (Table 3), and Study 5 (Table 5) in the article [8], respectively. In the Japanese and English experiments (Studies 1 and 2), a stimulus video showed an actor holding an object for around half a second, to ensure that children could clearly see the object. The actor then begins to act with the object. The object-holding segment was removed in the Chinese experiment (Study 5), because Chinese-speaking children tend to select the highlighted object in test trials. In Study 3, Chinese-speaking children were examined using the same video stimuli as the Japanese and English experiments, including the object-holding segment. The results demonstrated that a significant number of five-year-olds incorrectly considered a novel verb as the object. In contrast, Chinese five-year-olds were able to select an action as a novel verb in Study 5. When Japanese and English three-year-olds were tested using the video without the object-holding segment (Study 6), the results did not significantly differ from those of Studies 1 and 2. Imai *et al.* [8] discussed the difference between Studies 3 and 5, stating that Chinese children relied on a perceptual (extralinguistic) cue more than Japanese and English children did to distinguish between nouns and verbs, because Chinese language does not contain any morphologies to distinguish them.

Our model does not aim to explain the effect of the extralinguistic cue, but rather an effect of a linguistic (morphosyntactic) cue. Therefore, we used the result of Study 5 as the Chinese data. If possible, we should have used the result of Study 6, where the same stimuli was employed as in Study 5. However, only three-year-olds were examined in Study 6, and not five-year-olds. For this reason, we considered the results of Studies 1 and 2. We note again that it was confirmed that the results for three-year-old in Studies 1 and 2 did not significantly differ from those of Study 6. In general, younger children tend to use the perceptual cue. Therefore, it is speculated that there is not a significant difference between the stimuli with and without the object-holding segment for five-year-olds.

APPENDIX B
LEARNING METHODS

The model learns $P(s_t|w_t, \mathbf{s}_{-t})$ and $p(m|s_t)$ in (1). These probabilities represent the induction of a syntactic category s_t for a word w_t, and the association of the semantic category m with s_t, respectively. Although children are not obviously taught grammatical knowledge, e.g., part-of-speech categories, they can acquire syntax and lexicon knowledge from daily verbal communications. Therefore, the model bootstraps these probabilities from the semantic features in scenes and word sequences that explain the scenes without any teaching signals. It is assumed that the semantic features and words are segmented into discrete variables *a priori*.

A. Acquisition of Syntactic Categories

We apply an HMM to compute $P(s_t|w_t, \mathbf{s}_{-t})$. This is an unsupervised learning method, which determines a hidden Markov chain structure from some time-series input. When the input data items are given as sequential words (sentences), the acquired hidden states consequently represent the part-of-speech categories of the words, and the transition probabilities between the categories are interpreted as simple syntactic structures. HMMs are often employed as probabilistic models for assigning part-of-speech tags to words. They are trained using untagged text in natural language processing [10], [70]. The Bayesian HMM (BHMM) introduces Bayesian inference to the HMM, in order to improve the accuracy of tagging, especially in smaller learning corpora [34]. However, the basic functionality of these models is identical.

This model assumes that each s_t in a sequential hidden state $\mathbf{s} = (s_1, s_2, \ldots, s_t, \ldots)$ stochastically generates a word w_t, which produces a time-series of words $\mathbf{w} = (w_1, w_2, \ldots, w_t, \ldots)$. In this paper, we considered a trigram structure for $\mathbf{s}$, i.e., a hidden state s_t depends on the previous two states (s_{t-2} and s_{t-1}). The semantic features o and f is not used in the learning and inference of the hidden states in the BHMM, although Fig. 1 in the main text illustrates the probabilistic relations ($P(o|w_t)$ and $P(o|m)P(m|s_t)$) for inferring o.

Under this model, Gibbs sampling [71] estimates the posterior distributions of $\mathbf{s}$ and $\mathbf{w}$. Please see [34] for the detailed computation.

The corpus $\mathbf{w}$ used for the learning contained two dummy words, beginning of sentence (BOS) and end of sentence (EOS), to denote the beginning and the end of a sentence, respectively. For example, $\mathbf{w} =$ (BOS, she, is, read, -ing, a, book, EOS, BOS, the, chair, $\cdots$, EOS). We assigned values of 1 and 2 to the hidden states of BOS and EOS, respectively, and the probabilities of these states were not updated. The initial values of hidden states other than the dummy words were randomly allocated values between 3 and S. Note that in the main text, S does not include the two dummy word states. Then, we iteratively calculated the above probability 500 times, and averaged the latter 80% of samplings to obtain $P(s_t|\mathbf{w}, \mathbf{s}_{-t})$. In all experiments, the hyper-parameters were empirically set to $\alpha = 0.9$ and $\beta = 1.0$.

B. Learning of Linguistic-Feature Associations

After the induction of the syntactic categories of each word, the model learns to associate the linguistic categories ($\mathbf{w}$ and $\mathbf{s}$) with the semantic categories (f and m) from utterance-scene pairs. The ith utterance ($i = 1, 2, \ldots, N$) consists of a time-series words $\mathbf{w}^{(i)} = (w_1^i, w_2^{(i)}, \ldots, w_t^{(i)}, \ldots, w_{T^{(i)}}^{(i)})$, and their syntactic categories are parameterized as $\mathbf{s}^{(i)} = (s_1^{(i)}, s_2^{(i)}, \ldots, s_t^{(i)}, \ldots, s_{T^{(i)}}^{(i)})$. At the same time as the presence of the utterance, the model observes a set of semantic features $f^{(i)}$ in the ith scene. This set is not a time-series. The probability that a word w_q is associated with a semantic feature f_p is simply calculated using their co-occurrence probability

$$P\left(o = f_p | w = w_q\right) \equiv \frac{1}{Z} \sum_i^N n^{(i)}\left(w_q, f_p\right).$$

Here, $n^{(i)}(w_q, f_p)$ denotes the number of times that w_q and f_p appear simultaneously in the ith utterance-scene pair $(\mathbf{w}^{(i)}, f^{(i)})$, and Z is a normalizing constant. We assume that this probability can be applied to all word variables: $P(o|w_t^{(i)}) = P(o|w)$ ($t = 1, 2, \ldots, T^{(i)}$ and $i = 1, 2, \ldots, N$).

The probability of m under s_t, i.e., $P(m|s_t)$, is subsequently calculated. This probability is expected to indicate that an action category, for example, tends to correspond to a verb category. Such a mapping appears to be computed using the co-occurrence, in the same manner as the above equation. However, it is difficult to acquire meaningful mappings by calculating the simple co-occurrence, because the probability of m is practically a uniform distribution. Given the ith scene, the model observes a set of many semantic features $f^{(i)}$ that consist of objects and actions, resulting in an almost uniform distribution over the object and action categories in m. Therefore, the model cannot significantly associate s_t with m. To avoid this, we expand the semantic categories in the ith scene into a time-series of semantic categories $\mathbf{m}^{(i)} = (m_1^{(i)}, m_2^{(i)}, \ldots, m_t^{(i)}, \ldots)$, corresponding to the word series. The sequence of semantic categories is obtained using the well-learned $P(o|w)$

$$P\left(m_t^{(i)} | w_t^{(i)}\right) = \sum_o P\left(m_t^{(i)} | o\right) P\left(o | w_t^{(i)}\right).$$

TABLE I
VOCABULARY SIZES OF WORD CATEGORIES IN EACH CORPUS

	Subject	Object	Adjective	Transitive	Intransitive	Article	Auxiliary	Adverb
English	41	37	35	20	14	2	-	-
Japanese	45	45	34	21	15	-	3	-
Chinese	56	56	34	20	14	-	-	2

Here, the first term on the RHS (mappings from features to semantic categories) is known *a priori*, and the second term is acquired using the previous equation. $P(m|s)$ is then obtained by enumerating the co-occurrence between $m_t^{(i)}$ and $s_t^{(i)}$ over t and i

$$P\left(m = m_q | s = s_p\right) \equiv \frac{1}{Z} \sum_{i}^{N} \sum_{t}^{T^{(i)}} P\left(m_t^{(i)} = m_q | w_t^{(i)}\right) P\left(s_t^{(i)} = s_p | w_t^{(i)}, \mathbf{s}_{-t}^{(i)}\right)$$

where Z denotes a normalizing constant. This probability is assumed to be applicable to all syntactic variables, i.e., $P(m|s_t^{(i)}) = P(m|s)$ $(t = 1, 2, \ldots, T^{(i)}$ and $i = 1, 2, \ldots, N)$.

APPENDIX C
LEARNING CORPORA

The model was trained using a simple artificial corpus in English, Japanese, or Mandarin Chinese, consisting of nouns, verbs, adjectives, and so on. Fig. 7 illustrates the transition rules between word categories for sentence production for the (a) English, (b) Japanese, and (c) Mandarin Chinese corpora. Each circle in Fig. 7 indicates the word category from whose vocabulary a word is randomly produced. We assume that each word is unambiguously chosen from just one category. There is a 50% chance that an adjective category is added before a noun (the subject and object categories) in all of the corpora, although Fig. 7 does not depict the categories. Table I lists the size of the vocabulary for each word category. The vocabularies of the subject and object categories consists of the same nouns, but words in the nominative case are added to the vocabulary of the English subject category. These rules start at the BOS, and end at the EOS. The syntactic categories of these dummy words are allocated in advance, and are not used in the inference of the target semantic feature. The blue decimals above the arrows denote the transition probabilities between word categories. The orange decimals above the blue circles give the probabilities that the word categories are dropped off. In a sentence including a verb, there are three possible tenses, i.e., present, past, and progressive forms, which are selected with equal probability.

In the English rules [see Fig. 7(a)], a word from the article category is placed before words from the categories of nouns or adjectives. Although the nouns are only used in the singular, the verb suffixes "-s," "-es," or "-ies" are omitted in the present tense. However, the verb suffixes "-ed" and "-ing" can be arranged after verbs to represent the past and progressive forms, respectively. In the Japanese rules [see Fig. 7(b)], the verb suffixes (auxiliary) determine the tense. There are three particles in Japanese. These indicate the semantic roles (subject or object) of the nouns that precede and follow the particles. Particle1 and particle3 denote the nominal particles "ga" and "ha," respectively, and their preceding nouns are subject words. Particle2 is the accusative particle "wo," indicating that the preceding noun is an object word. We separate the adjective suffix "i" from adjective stems, because most

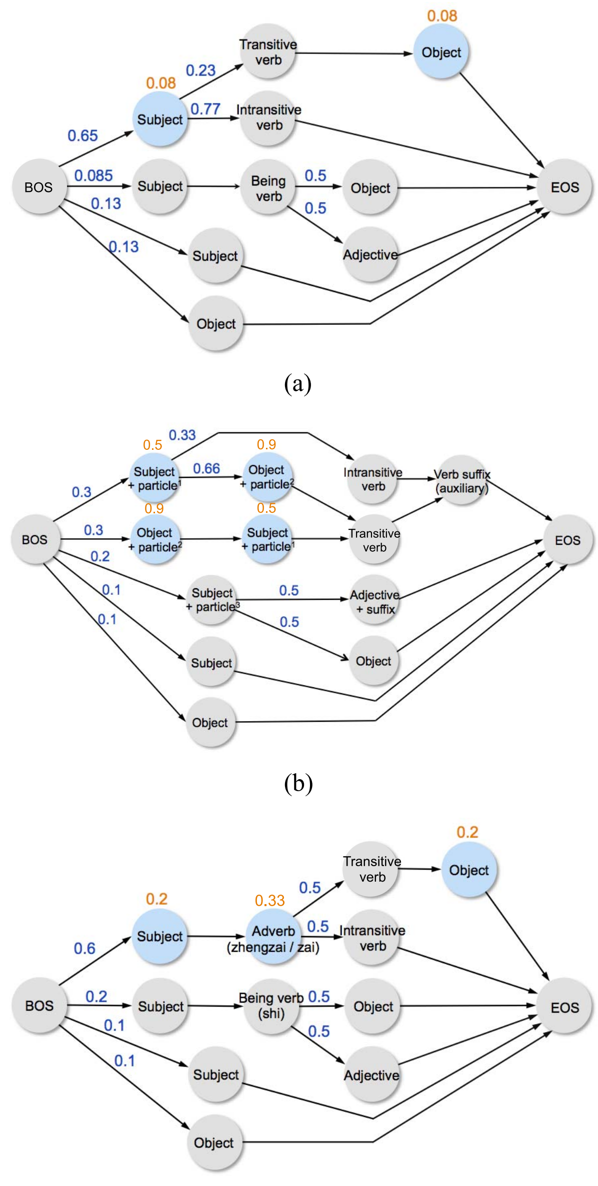


Fig. 7. Transition rules between word categories for the production of the corpora. These figures show rules for generation of the (a) English, (b) Japanese, and (c) Chinese utterances. The blue and orange decimals indicate the transition probabilities between word categories and the probabilities that a word category is dropped off, respectively. There is a 50% chance that an adjective category is placed before the subject and object categories in all corpora.

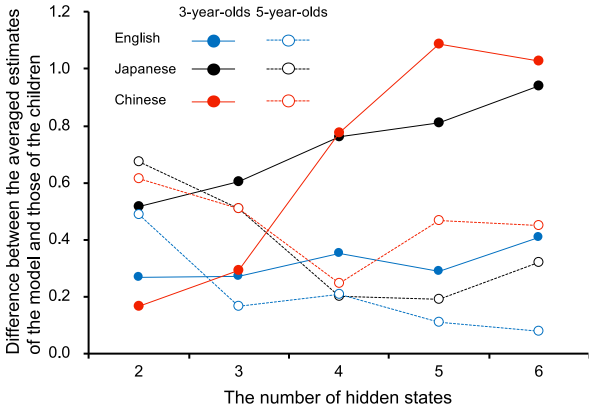


Fig. 8. Fitness between model performance and that of the children.

Japanese adjectives have this suffix. In the Mandarin rules [see Fig. 7(c)], the adverb ("zhengzai" or "zai") represents the progressive form. The "shi" is used as a be-verb in copulas (sentences in which the verb is some form of "to be"), although it actually has additional grammatical roles. In this paper, the particle "de" is only employed as the adjective suffix, which is arranged following adjective stems.

The transition probabilities are determined such that certain syntactic characteristics of the artificial corpora fit those of the real corpora of child-directed utterances, i.e., English Belfast Corpus [36], [72], Japanese MiiPro Nanami Corpus [37], [73], and Chinese Beijing 2 Corpus [38], [74]. These corpora have speech data of four, eight, and ten children, respectively. Each of their total word length is approximately hundred thousand. We calculated the percentages with which predicate verbs, subjects, objects, single sentences consisting a noun, and bare verb sentences occur in the real corpora. We then ensured that these percentages were almost identical in the artificial corpora with the real ones, by setting appropriate transition probabilities. In addition, the English corpus has the same ratio of three construction types: 1) fragments (sentences without subject and predicate); 2) copulas; and 3) sentences containing both a subject and a single predicative intransitive, as in the real child-directed speech reported by Cameron-Faulkner *et al.* [40].

Appendix D
Fitness Between Model and Children

We evaluated the absolute differences between the averaged model performance and that of the children, i.e., the differences between bars and dots in Fig. 4. Fig. 8 shows the differences for different numbers of categories. In the case of three-year-olds (the filled circles), the model with two hidden states exhibited the best fitness for the children's behavior in all three languages. In the case of five-year-olds (the empty circles), the adequate numbers of hidden sates differed depending on languages: six, five, and four states fitted behavioral data of English, Japanese, and Chinese children, respectively.

Acknowledgment

The authors gratefully acknowledge the advice of Prof. M. Imai of Keio University.

References

[1] M. D. S. Braine and M. Bowerman, "Children's first word combinations," *Monographs Soc. Res. Child Develop.*, vol. 41, no. 1, pp. 1–104, 1976.
[2] R. Olguin and M. Tomasello, "Twenty-five-month-old children do not have a grammatical category of verb," *Cogn. Develop.*, vol. 8, no. 3, pp. 245–272, 1993.
[3] M. Tomasello and P. J. Brooks, "Young children's earliest transitive and intransitive constructions," *Cogn. Linguist.*, vol. 4, no. 9, pp. 370–395, 1998.
[4] M. Tomasello, "Do young children have adult syntactic competence?" *Cognition*, vol. 74, no. 3, pp. 209–253, 2000.
[5] M. Tomasello, "The item-based nature of children's early syntactic development," *Trends Cogn. Sci.*, vol. 4, no. 4, pp. 156–163, 2000.
[6] M. Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, U.K.: Harvard Univ. Press, 2003.
[7] M. Dittmar, K. Abbot-Smith, E. Lieven, and M. Tomasello, "Children aged 2 ; 1 use transitive syntax to make a semantic-role interpretation in a pointing task," *J. Child Lang.*, vol. 38, no. 5, pp. 1109–1123, 2011.
[8] M. Imai *et al.*, "Novel noun and verb learning in Chinese-, English-, and Japanese-speaking children," *Child Develop.*, vol. 79, no. 4, pp. 979–1000, 2008.
[9] M. Imai, E. Haryu, and H. Okada, "Mapping novel nouns and verbs onto dynamic action events: Are verb meanings easier to learn than noun meanings for Japanese children?" *Child Develop.*, vol. 76, no. 2, pp. 340–355, 2005.
[10] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
[11] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.
[12] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based *n*-gram models of natural language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, 1992.
[13] S. Finch and N. Chater, "Bootstrapping syntactic categories using statistical methods," in *Background and Experiments in Machine Learning of Natural Language*, W. Daelemans and D. Powers, Eds. Tilburg, The Netherlands: Tilburg Univ., 1992, pp. 229–235.
[14] H. Schütze, "Part-of-speech induction from scratch," in *Proc. 31st Annu. Meeting Assoc. Comput. Linguist.*, 1993, pp. 251–258.
[15] M. Redington, N. Chater, and S. Finch, "Distributional information: A powerful cue for acquiring syntactic categories," *Cogn. Sci.*, vol. 22, no. 4, pp. 425–469, 1998.
[16] A. Clark, "Inducing syntactic categories by context distribution clustering," in *Proc. 2nd Workshop Learn. Lang. Logic 4th Conf. Comput. Nat. Lang. Learn.*, 2000, pp. 91–94.
[17] T. H. Mintz, "Frequent frames as a cue for grammatical categories in child directed speech," *Cognition*, vol. 90, no. 1, pp. 91–117, 2003.
[18] C. Parisien, A. Fazly, and S. Stevenson, "An incremental Bayesian model for learning syntactic categories," in *Proc. 12th Conf. Comput. Nat. Lang. Learn.*, 2008, pp. 89–96.
[19] A. Alishahi and G. Chrupała, "Lexical category acquisition as an incremental process," in *Proc. CogSci Workshop Psychocomput. Mod. Human Lang. Acquisition*, 2009.
[20] L. Onnis and M. H. Christiansen, "Lexical categories at the edge of the word," *Cogn. Sci.*, vol. 32, no. 1, pp. 184–221, 2008.
[21] L. M. Santelmann and P. W. Jusczyk, "Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space," *Cognition*, vol. 69, no. 2, pp. 105–134, 1998.
[22] T. H. Mintz, "The segmentation of nlexical morphemes in English-learning 15-month-olds," *Front. Psychol.*, vol. 4, p. 24, Feb. 2013.
[23] P. Monaghan, M. H. Christiansen, and N. Chater, "The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition," *Cogn. Psychol.*, vol. 55, no. 4, pp. 259–305, 2007.
[24] F. T. Asr, A. Fazly, and Z. Azimifar, "The effect of word-internal properties on syntactic categorization: A computational modeling approach," in *Proc. 32nd Annu. Conf. Cogn. Sci. Soc.*, 2010, pp. 1529–1535.
[25] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, nos. 1–2, pp. 39–91, 1996.
[26] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558–1568, 2008.
[27] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "A Bayesian framework for cross-situational word-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 457–464.

[28] A. Fazly, A. Alishahi, and S. Stevenson, "A probabilistic computational model of cross-situational word learning," *Cogn. Sci.*, vol. 34, no. 6, pp. 1017–1063, 2010.

[29] A. Alishahi and A. Fazly, "Integrating syntactic knowledge into a model of cross-situational word learning," in *Proc. 32nd Annu. Conf. Cogn. Sci. Soc.*, 2010, pp. 2452–2457.

[30] A. Toyomura and T. Omori, "A computational model for taxonomy-based word learning inspired by infant developmental word acquisition," *IEICE Trans. Inf. Syst.*, vol. 88, no. 10, pp. 2389–2398, 2005.

[31] C. Yu, "Learning syntax-semantics mappings to bootstrap word learning," in *Proc. 28th Annu. Conf. Cogn. Sci. Soc.*, 2006, pp. 924–929.

[32] A. Alishahi and G. Chrupała, "Concurrent acquisition of word meaning and lexical categories," in *Proc. Joint Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, 2012, pp. 643–654.

[33] M. Attamimi *et al.*, "Learning word meanings and grammar for verbalization of daily life activities using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models," *Adv. Robot.*, vol. 30, nos. 11–12, pp. 806–824, 2016.

[34] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, vol. 45, 2007, pp. 744–751.

[35] E. M. Markman and G. F. Wachtel, "Children's use of mutual exclusivity to constrain the meanings of words," *Cogn. Psychol.*, vol. 20, no. 2, pp. 121–157, 1988.

[36] A. Henry, *English Belfast Corpus*. Pittsburgh, PA, USA: TalkBank, 2004.

[37] S. Miyata, *MiiPro Nanami Corpus*. Pittsburgh, PA, USA: TalkBank, 2013.

[38] T. Tardif, *Chinese Beijing Corpus*. Pittsburgh, PA, USA: TalkBank, 2007.

[39] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk: Vol. II: The Database*, 3rd ed. Mahwah, NJ, USA: Lawrence Erlbaum Assoc., 2000.

[40] T. Cameron-Faulkner, E. Lieven, and M. Tomasello, "A construction based analysis of child directed speech," *Cogn. Sci.*, vol. 27, no. 6, pp. 843–873, 2003.

[41] M. S. Seidenberg, "Language acquisition and use: Learning and applying probabilistic constraints," *Science*, vol. 275, no. 5306, pp. 1599–1603, 1997.

[42] M. S. Seidenberg and M. C. MacDonald, "A probabilistic constraints approach to language acquisition and processing," *Cogn. Sci.*, vol. 23, no. 4, pp. 569–588, 1999.

[43] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *Int. J. Robot. Res.*, vol. 23, nos. 4–5, pp. 363–377, 2004.

[44] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[45] D. Klein and C. D. Manning, "A generative constituent-context model for improved grammar induction," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 128–135.

[46] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, "Grounding of word meanings in latent Dirichlet allocation-based multimodal concepts," *Adv. Robot.*, vol. 25, no. 17, pp. 2189–2206, 2011.

[47] E. L. Newport, "Maturational constraints on language learning," *Cogn. Sci.*, vol. 14, no. 1, pp. 11–28, 1990.

[48] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, 1993.

[49] K. L. Sakai, "Language acquisition and brain development," *Science*, vol. 310, no. 5749, pp. 815–819, 2005.

[50] P. M. Thompson *et al.*, "Growth patterns in the developing brain detected by using continuum mechanical tensor maps," *Nature*, vol. 404, no. 6774, pp. 190–193, 2000.

[51] J. Matsuzawa *et al.*, "Age-related volumetric changes of brain gray and white matter in healthy infants and children," *Cerebral Cortex*, vol. 11, no. 4, pp. 335–342, 2001.

[52] A. L. Theakston, E. V. M. Lieven, J. M. Pine, and C. F. Rowland, "The role of performance limitations in the acquisition of verb-argument structure: An alternative account," *J. Child Lang.*, vol. 28, no. 1, pp. 127–152, 2001.

[53] S. Choi and A. Gopnik, "Early acquisition of verbs in Korean: A cross-linguistic study," *J. Child Lang.*, vol. 22, no. 3, pp. 497–529, 1995.

[54] C. E. Snow, "Mothers' speech to children learning language," *Child Develop.*, vol. 43, no. 2, pp. 549–565, 1972.

[55] J. R. Phillips, "Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons," *Child Develop.*, vol. 44, no. 1, pp. 182–185, 1973.

[56] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 577–584.

[57] J. Van Gael, A. Vlachos, and Z. Ghahramani, "The infinite HMM for unsupervised PoS tagging," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2009, pp. 678–687.

[58] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2012.

[59] D. Gentner, *Why Nouns Are Learned Before Verbs: Linguistic Relativity Versus Natural Partitioning*, S. A. Kuczaj, Ed., Erlbaum, Hillsdale, 1982.

[60] D. Gentner and L. Boroditsky, "Lndividuation, relativity, and early word learning," in *Language Acquisition and Conceptual Development*, M. Bowerman and S. C. Levinson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2001, pp. 215–256.

[61] M. H. Bornstein *et al.*, "Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English," *Child Develop.*, vol. 75, no. 4, pp. 1115–1139, 2004.

[62] T. Ogura, P. S. Dale, Y. Yamashita, T. Murase, and A. Mahieu, "The use of nouns and verbs by Japanese children and their caregivers in book-reading and toy-playing contexts," *J. Child Lang.*, vol. 33, no. 1, pp. 1–29, 2006.

[63] M. J. Maguire, K. Hirsh-Pasek, and R. M. Golinkoff, "A unified theory of word learning: Putting verb acquisition in context," in *Action Meets Word: How Children Learn Verbs*, K. Hirsh-Pasek and R. M. Golinkoff, Eds. New York, NY, USA: Oxford Univ. Press, 2006, pp. 364–391.

[64] P. Zukow-Goldring, "Socio-perceptual bases for the emergence of language: An alternative to innatist approaches," *Develop. Psychobiol.*, vol. 23, no. 7, pp. 705–726, 1990.

[65] P. Zukow-Goldring, *Perceiving Referring Actions: Latino and Euro-American Infants and Caregivers Comprehending Speech*, vol. 11. Mahwah, NJ, USA: Lawrence Erlbaum Assoc., Inc., 2001, pp. 139–165.

[66] L. J. Gogate, L. H. Bolzani, and E. A. Betancourt, "Attention to maternal multimodal naming by 6-to 8-month-old infants and learning of word–object relations," *Infancy*, vol. 9, no. 3, pp. 259–288, 2006.

[67] K. J. Rohlfing, B. Wrede, A.-L. Vollmer, and P.-Y. Oudeyer, "An alternative to mapping a word onto a concept in language acquisition: Pragmatic frames," *Front. Psychol.*, vol. 7, p. 470, Apr. 2016.

[68] B. MacWhinney, Ed., "The competition model," in *Mechanisms of Language Acquisition*. Hillsdale, NJ, USA: Erlbaum, 1987, pp. 249–308.

[69] E. Bates and B. MacWhinney, Eds., "Functionalism and the competition model," in *The Crosslinguistic Study of Sentence Processing*. Cambridge, U.K.: Cambridge Univ. Press, 1989, pp. 3–73.

[70] B. Merialdo, "Tagging English text with a probabilistic model," *Comput. Linguist.*, vol. 20, no. 2, pp. 155–171, 1994.

[71] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[72] A. Henry, Ed., *Belfast English and Standard English: Dialect Variation and Parameter Setting*. New York, NY, USA: Oxford Univ. Press, 1995.

[73] Y. Oshima-Takane and B. MacWhinney, Eds., *CHILDES Manual for Japanese*. Montreal, QC, Canada: McGill Univ., 1995.

[74] T. Tardif, "Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies," *Develop. Psychol.*, vol. 32, no. 3, pp. 492–504, 1996.

Yuji Kawai received the M.S. and Ph.D. degrees in engineering from Osaka University, Suita, Japan, in 2013 and 2017, respectively.

He has been an Assistant Professor at the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University since 2017. His current research interests include constructive understanding of human cognitive development and computational neuroscience.

Yuji Oshima received the M.S. degree in engineering from Osaka University, Suita, Japan, in 2014.

He has been a Researcher with NTT Software Innovation Center, Tokyo, Japan, since 2014. His current research interests include natural language processing and machine learning.

Yuki Sasamoto received the M.S. and Ph.D. degrees in engineering from Osaka University, Suita, Japan, in 2010 and 2015, respectively.

He was a Researcher at Fujitsu Laboratories Ltd., Kawasaki, Japan, from 2014 to 2017 and Fujitsu Laboratories of Europe Ltd., London, U.K., from 2017 to 2018. He has been a Senior Customer Solutions Architect at Fujitsu Finland Oy, Helsinki, Finland, since 2018. His current research interest includes human sensing and its business application in healthcare.

Yukie Nagai (M'09) received the M.S. degree in engineering from Aoyama Gakuin University, Tokyo, Japan, in 1999 and the Ph.D. degree in engineering from Osaka University, Suita, Japan, in 2004.

She was a Postdoctoral Researcher at the National Institute of Information and Communications Technology (NICT), Kyoto, Japan, from 2004 to 2006, and Bielefeld University, Bielefeld, Germany, from 2006 to 2009. She was a Specially Appointed Associate Professor at Osaka University from 2009 to 2017, and became a Senior Researcher at NICT in 2017. She was a Project Leader of MEXT Grant-in-Aid for Scientific Research on Innovative Areas Computational Modeling of Social Cognitive Development and Design of Assistance Systems for Developmental Disorders from 2012 to 2017, and has been the Project Leader of JST CREST Cognitive Mirroring: Assisting People With Developmental Disorders by Means of Self Understanding and Social Sharing of Cognitive Processes since 2016. Her current research interest includes computational modeling of human cognitive functions, such as self-other cognition, imitation and joint attention, and design of assistant systems for developmental disorders.

Minoru Asada (F'05) received the B.E., M.E., and Ph.D. degrees in control engineering from Osaka University, Osaka, Japan, in 1977, 1979, and 1982, respectively.

In 1995, he became a Professor at Osaka University, where he has been a Professor at the Department of Adaptive Machine Systems, Graduate School of Engineering since 1997.

Dr. Asada was a recipient of the Best Paper Award at the IEEE/RSJ International Conference on Intelligent Robots and Systems and a Commendation by the Minister of Education, Culture, Sports, Science and Technology, Japanese Government, as a Person of Distinguished Services to Enlightening People on Science and Technology. He is one of the founders of the RoboCup, and the Former President of the International RoboCup Federation from 2002 to 2008. He was the Research Director of the ASADA Synergistic Intelligence Project at Exploratory Research for Advanced Technology by the Japan Science and Technology Agency from 2005 to 2011, and was a Principal Investigator of the Grants-in-Aid for Scientific Research (Research Project Number: 24000012) from 2012 to 2016. He is currently a Principal Investigator of the JST RISTEX Research and Development Project titled "Legal beings: Electric personhoods of artificial intelligence and robots" in NAJIMI Society, based on a reconsideration of the concept of autonomy.