Cognitive model of phonological awareness focusing on errors and formation process through Shiritori

Jumpei Nishikawa and Junya Morita

Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan

ABSTRACT
Language acquisition is supported by phonological awareness, which intentionally makes children aware of phonological units. By understanding the internal processes of children during language acquisition, this study aims to elucidate factors that can correct erroneous phonological generation. Therefore, we developed a cognitive model using innate and experiential factors of the memory retrieval in the cognitive architecture—ACT-R. Furthermore, we performed simulations using Shiritori, a Japanese word game, as an interaction task. The simulation included the observation of effects of the experiential factor of repeating a task and innate factors of different settings. It showed that repeating a single task causes incorrect convergence, and this convergence can be prevented by comprehensive activation of overall phonological knowledge during the interval of Shiritori tasks. Moreover, the simulation in specific innate settings exhibited commonalities with cases of developmental disorder by showing errors like consonant deletion. In the future, we will examine the correlation of the aforementioned findings with actual language development to realize the use of cognitive architecture in real world.

ARTICLE HISTORY
Received 30 April 2021
Revised 17 August 2021 and 8 December 2021
Accepted 18 December 2021

KEYWORDS
Cognitive modeling; phonological awareness; mora; ACT-R

1. Introduction
The research topic ‘Symbol Emergence in Robotics (SER)’ aims to achieve natural man–machine interactions that symbols grounded in physical and semiotic environments mediate [1]. In the SER approach, these symbols are assumed to be constructed by a self-organizing process where segments extracted from time-series data are assigned meanings and not directly provided by humans. Although the endeavor comprised experimental developments of robotic systems acting in physical environments, it would be beneficial to achieve the goal of SER to computationally describe the process of symbol emergence in natural cognitive systems (human). Therefore, this study presents a simulation analysis of human language development, mainly oral language development.

The basic assumption of our simulation is that human language development is an experiential process that initiates from an innate basis, which is widely common among modern humans. This view follows literatures that relate to the traditional generative grammar [2] and emphasizes the role of learning and training in language acquisition [3]. We have vocal and aural organs and use primitive cognitive functions, e.g. joint attention [4] or role reversal imitation [5]. These innate factors help children acquire language-specific structures through experience made in their linguistic community.

Among several aspects of language ability, phonological segmentation is the most prominent example of showing such an experiential process based on innate bias. This aspect of language is treated in research with regard to symbol emergence in biological [6] and artificial systems [7]. In human language acquisition, infants initially have inborn substrates, making them possible to acquire various languages that separate sounds into different types of segments ([8] as a recent experimental study). The basis to acquire such units converges into a system to process units defined by mother tongue as a child interacts with adults in her/his community. This developmental process is partly attributed to phonological awareness ([9] as a review), which enables an individual to focus on phonological aspects of oral language, e.g. phonemes and rhythms.

The involvement of innate bases for phonological acquisitions is evident when we see cases of genetic developmental disorders, e.g. autism spectrum disorder (ASD). Not a few children with ASD have trouble acquiring language-specific phonological structure of
their mother tongue. To support those who suffer from such innate difficulty, speech-language pathologists, who target the early developmental process, focus on phonological awareness as a trainable cognitive ability.

Researchers and practitioners in the field often report phonological errors and work out individual cases of such errors [10, 11]. Thus, the role of phonological awareness and conditions that phonological awareness is formed have been identified based on experimental and clinical field research. However, the cognitive process by which an innate basis leads to erroneous phonological states and recovers through proper phonological awareness is not straightforward. To clarify this internal process, the present study develops a computational model that represents

(1) development process that leads to erroneous phonological generation and
(2) experiential factors involved in the suppression of phonological errors.

We consider that achieving these goals contributes to the support for human language disorders and the development of the SER technique because the technique explicitly shows hypothetical symbol emergent processes from various innate bases. For this motivation, we use basic cognitive functions implemented in the cognitive architecture – Adaptive Control of Thought – Rational (ACT-R [12]). Before presenting our model, we will provide the background of our study.

2. Related research

To build this study, we present a framework of representing sounds used in oral language. Following this framework, we introduce studies of phonological awareness as a reference of our targeting phenomena. Finally, we show previous work in cognitive modeling on language development to introduce our method of modeling.

2.1. Representation of phonological elements

To handle phonological segmentation in a model, we need to formally represent oral languages. In phonology, researchers have developed systems to feature sounds that humans perceive and utter. The most famous system is the distinctive feature [13]. It represents features that have binary variables, with a value of + or −, classified based on movements of the tongue and throat associated with an utterance. Table 1 lists definitions of distinctive features for some phonemes, which are used in the later section (taken from [13]). In the table, while the cells filled with + and − indicate whether the phoneme includes the feature or not, the blank cells obscure the existence of the feature in the phoneme. In actual utterance, the values in blank cells change depending on the other phonemes of the words.

These units featuring biological vocal tract organs were originally defined to represent universal sound patterns, which did not depend on a specific language [13]. In other words, the combinations between the units and the features in the table are assumed as the innate endowment that is available for children who learn their own language [13]. However, there are various combinatorial systems for those units, i.e. languages, worldwide. In other words, basic sound units of various languages are different. According to Trubetzkoy [17], such systems fall into groups based on the prosodic unit commonly used in that language. One of them is the ‘Mora language,’ whose basic unit is mora, which defines a unit of duration [18]. In Japanese, one of the mora languages, morae consist of five vowels (a, i, u, e, o), 59 vowel-consonant combinations (ka, ki, ku, ke, ko, sa, si, and so forth), and some special mora. Phonemes used in Japanese morae are listed in Table 1.

2.2. Studies on phonological awareness

To describe the development process of phonological segmentation, this study focuses on phonological awareness. Although numerous definitions have been offered, one of the literatures noted it as ‘one’s degree of sensitivity to the sound structure of oral language’ [19]. Moreover, the same literature stated that this ability is measured by several tasks of blending sounds together, segmenting words into their constituent sounds, recombining sounds of words, and judging whether two words have some sounds in common. It has been clarified that the performance of these tasks with regard to preschool children relates to the future acquisition of reading literacy [9]. Based on these studies, it can be considered that phonological awareness is assumed to be the ability making the correspondence of verbalized sound pattern and internal sound representation, which often corresponds to graphical characters, e.g. alphabets.

Concerning this definition of phonological awareness, Japanese has an advantage for research because correspondence between sound patterns and graphical characters is clear. With few exceptions, one mora in Japanese essentially corresponds to one kana (Japanese character). We think that this characteristic is beneficial for modeling phonological awareness with regard to a symbolic cognitive architecture. Therefore, in this study, we focus on Japanese mora development as the target of modeling.

Several clinical or educational studies in Japanese report that children mispronounce certain sounds. For
example, it is reported that two- or three-year-old children tend to pronounce ‘r’ as ‘d.’ This mispronunciation error reduces by the time they reach the age of four or five [20]. On the other hand, there are cases of delay in such phonological discrimination. In the text of Japanese speech-language pathology, it has been reported that children who have a developmental disorder have difficulty in distinguishing a vowel and a vowel-consonant combination that holds the same vowel (e.g. ‘a’ and ‘ka’) [21].

To examine the development of phonological awareness suppressing such errors, several Japanese studies have used Shiritori – a popular word game – as a task. In Shiritori, players turn turns uttering a word (noun); the word must begin with the mora that the previous word ended with. For example, after a player answers ri-n-go (apple), the next player continues with go-ri-ra (gorilla). The game is over when a player provides a word that ends with ‘N,’ because no Japanese word can begin with this special mora. For example, go-ri-ra → ra-i-ne-n (next year) → game over. To avoid looping, the game ends if a player repeats a word that has already been used as an answer in the game: ma-su-ku (mask) → ku-ru-ma (car) → ma-su-ku → game over.

Takahashi [22] examined conditions required to play Shiritori and the stages of phonological awareness formation based on a cross-sectional developmental experiment with typically developing children. This research indicated that playing Shiritori requires being able to divide sounds into morae and a mental lexicon indexed by morae. Moreover, it indicated that the acquisition of kana characters is effective to index vocabulary based on morae. In other words, playing Shiritori requires phonological awareness that focuses on morae. Moreover, this study showed that children who do not have a proper level of phonological awareness required for Shiritori can participate in Shiritori with adult assistance.

Based on these findings, the current study focuses on the ability of phonological manipulation, such as extracting an ending mora from a word and retrieving a word by its initial mora. Moreover, we use Shiritori as a task, with reference to Takahashi [22].

### 2.3. Cognitive modeling

Cognitive modeling is a scientific approach to constructing a computational approximation of human cognitive processes in a specific task. Several studies have been conducted to model the development process of phonological elements using artificial neural networks (ANNs). In the classic work by Rumelhart and McClelland [23], English phoneme sequences were deployed to an ANN, which produced irregular verb changes based on phoneme similarity. In a recent study, Matusevych [24] showed that inputting a multilingual speech corpus into an autoencoder enabled language-specific phonological structures to be extracted. Most of these connectionist studies do not make a priori assumptions about mapping between the structure of the network and functions. Many network structures are general and are not specific to the language domain. These characteristics are advantageous to develop efficient industrial products, though they make it difficult to incorporate known innate factors into the structure of the model based on a top–down approach. In addition, most studies based on ANN encounter a black-box problem, which implies that the learned structure is difficult to explain. Thus, these studies lack an explicit relationship with the existing findings discussed in phonology, as described in the previous section.

In contrast to such a simple connectionism approach, studies have been conducted using cognitive architecture for cognitive modeling. Cognitive architecture is a design specification that integrates intelligence-generating structures (brain structures) and intellectual functions (e.g. thought processes with regard to specific tasks) [12]. It can be regarded as the foundation (structure) of cognitive modeling, which accumulates cognitive
functions used in various tasks. A model using cognitive architecture allows us to isolate factors involved in accomplishing a task using a general structure.

Based on various cognitive architectures that have been developed, we have selected ACT-R [12] for this study. As ACT-R is based on psychological experiments with regard to thought process and memory, it enables us to comprehensively grasp various phenomena related to human cognition. The ACT-R structure is represented as a production system with multiple modules, where the central production module controls other modules. In addition, various parameters specify the behavior of the modules, facilitating the modeling of varieties of individuals. The current study focuses on the knowledge representation that can be expressed as discrete symbols provided by the modeler. It would be useful to consider how to map continuous sounds with phonological units (i.e. symbols) based on the knowledge representation in ACT-R to explain phonological awareness through modeling.

Several studies on language acquisition using ACT-R have been conducted. For instance, a model that acquires irregular English verbs [25] and a model of infants who are learning nouns [26] have been developed. In addition, brain dysfunction has been modeled based on ACT-R, and some studies have explained errors in sentence comprehension in aphasia using ACT-R parameters [27]. Among them, Nishikawa and Morita [28] built a model that maps phonological awareness to the knowledge activation mechanism of ACT-R. They showed that sound similarity is associated with innate factors in language development, and partial match retrieval based on that similarity can explain errors in the phonological awareness formation process. Moreover, they showed that manipulating parameters that correspond to the effect of learning associated with experiential factors can reduce the relative importance of similarity and suppress errors. However, the previous study does not explain how such experiential factors are developed. In the current study, we extended the previous model to phonological awareness formation and error suppression based on the task process.

3. Model

Based on the aforementioned research, we have developed a model to represent errors that occur in the development of phonological awareness and factors that suppress them. To achieve this, we use Shiritori (shown in the previous study) as a task. Moreover, we use the ACT-R architecture to study factors that contribute to the occurrence and suppression of errors. The ACT-R architecture has two main components: symbolic and subsymbolic. In a symbolic part, symbols implemented by a modeler describe the process and structure of the target task. In a subsymbolic part, these symbols are connected with the continuous world. This section describes the proposed model according to each component as the symbolic module structure and subsymbolic parameters for knowledge retrieval.

3.1. Module structure

The model is shown in Figure 1. Two agents (dashed line area) take turns answering with words to play Shiritori. One agent corresponds an adult, e.g. a caregiver, and the other agent represents a child, who is the subject of development. Each box of the agents corresponds to an ACT-R module. The interaction of those modules is mediated with a unit of symbols known as chunk. Each module includes a buffer temporarily holding chunks and controlled through the central production system. Next, we describe how the task of Shiritori is performed using these modules.

3.1.1. Declarative and imaginal modules

The declarative module plays the role of a database containing chunks required to execute the task. To execute the task, our model has three types of chunks, corresponding to word knowledge that connects pronunciation (sound pattern) and spelling of each word, mora knowledge that represents pronunciation and spelling of each mora, and knowledge of word–mora relationships that indicate what morae consist of the word. Table 2 lists examples of these chunks, where pronunciation and Kana spelling are described as double quoted letters in the form of International Phonetic Alphabet (IPA) and symbols formatted in Hepburn romanization, respectively.

In declarative memory, these three types of chunks are connected as a network. The balloon attached to the declarative modules shown in Figure 1 indicates the aforementioned structure. By tracing the network, the agents search for an answer word connecting to the previous word via a mora.

In addition, the model retains chunks about words that have already been answered during Shiritori. These chunks do not exist at the beginning of the trial, but they are generated as the trial progresses. In ACT-R, new chunks are created by a function of an imaginal module that is typically used to maintain context relevant to the current task [29]. After storing multiple chunks in the imaginal module, they are stored in the declarative module. In the model shown in Figure 1, chunks are generated by combining the knowledge of the answer word and the tag that indicates it has been played (e.g. ‘ri-n-go played’).
Table 2. Model’s declarative memory.

| Word   | Sound  |
|--------|--------|
| ringo  | ‘ringo’ |
| gorira | ‘gorira’|
| kuri   | ‘kuri’  |
| …      | …      |

(b) Phonological knowledge

| Mora  | Sound  |
|-------|--------|
| ri    | ‘rj’   |
| go    | ‘go’   |
| ku    | ‘kÉŕ’ |
| …    | …    |

(c) Word–mora relationship

| Mora   | Mora   | Position |
|--------|--------|----------|
| ri-n-go| ri     | head     |
| ri-n-go| go     | tail     |
| go-ri-ra| go     | head |
| …    | …    | …        |

3.1.2. Audio and speech modules

The ACT-R speech and audio modules simulate interactions of internal and external environments mediated with chunks. The speech module receives chunks from the production module and generates sound patterns held by the chunks to the external world, whereas the audio module receives sound patterns from the external world and transfers them to the production module. In the proposed model, the sound pattern output of the speech module of one agent becomes the input to the audio module of another agent. In this process, sound patterns exchanged by the two agents are stored in word knowledge (Table 2(a)). With regard to the sound of each word received by the audio module, the phonological structure of the word is interpreted by recalling the knowledge of the word–mora relationship (Table 2(c)). This process is controlled by the following goal and production modules.

3.1.3. Goal and production modules

The ACT-R production module selects rules that reference the current states of other modules and applies the selected rules to change the module states. Among these modules mediated by the production module, the goal module maintains the explicit control flow of the model [12].

Figure 2 shows the flow of the model. Blocks in the figure show the states maintained by the goal module, and their transitions are accomplished by production rules. First, the agent uses the auditory module to perceive the word that the other agent has answered (recognize the word). In this process, the sound pattern in the audio module is recognized by requesting a retrieval of the word knowledge (Table 2(a)) to the declarative module. Next, the production module requests the chunk that connects the word and ending mora (Table 2(c)) (retrieve the ending mora). As a result of this process, the mora knowledge (Table 2(b)) corresponding to the ending mora is obtained (gather the mora knowledge). Thus, the model perceives a word and segments it into a mora represented as chunks.

Then, based on this mora knowledge, the word that begins with the mora is retrieved (retrieve answer candidate), and the selected word is stored in the goal module as a candidate answer (get the answer candidate word).

After this, the model verifies that the stored candidate is valid according to the rules of Shiritori, i.e. not ending in ‘n’ (ending ‘N’?) and not having been previously answered in the current trial game (already answered?). If the current candidate violates these rules, the model searches again for a candidate answer (retrieve the initial mora of the answer candidate). When the candidate word is confirmed as valid (i.e. it fails to find a chunk of words connected with ‘n’ or a word that has already been answered), the model stores it in the declarative module as an answered word (annotate as already answered) and outputs the word through the speech module. During one task trial of Shiritori, the two agents alternately execute the above series of procedures until the given time limit is reached.
3.2. ACT-R parameters for knowledge retrieval

In the aforementioned process, phonological awareness, which is defined as an ability to focus on the mora, is involved in the retrieval of chunks of word–mora relationships (Table 2c). Such knowledge retrieval in ACT-R is controlled by an activation parameter that affects the success or failure of the retrieval. Thus, for ACT-R, it has been assumed that activation plays the role of attention (activated chunks are more likely to be aware). Therefore, phonological awareness and these errors can be modeled as the assignment of activation to proper and erroneous chunks.

In the standard ACT-R setting [29], the activation value of chunk $i$ is expressed as an addition of several terms:

$$A_i = B_i + S_i + P_i + \varepsilon_i$$  

where $B_i$ indicates the base level that represents the repetition and forgetting effects, $S_i$ is the contextual influence, $P_i$ represents the similarity of the knowledge, and $\varepsilon_i$ is a noise. When multiple chunks match the retrieval request from the production module, the chunk with the highest activation is selected. In this model, among the elements of activation, we focus on $P_i$ and $B_i$.

### 3.2.1. Similarity

As discussed in the former section, innate factors in the phonological development can be assumed as distinctive features [13]. In this study, we incorporated this innate bias into the knowledge similarity among mora chunks. $P_i$ of the activation assigned to chunk $i$ is computed as

$$P_i = PM_i$$  

where $P$ indicates the weight and $M_i$ represents the similarity between the chunk $i$ (the retrieval candidate stored in the declarative module) and the retrieval request by the production module. In ACT-R, $M_i$ ranges from 0 to $-1$, where 0 indicates that the chunk $i$ matches the current retrieval request. Thus, the closer $M_i$ tends to $-1$, the less match the retrieval candidate $i$ is to the retrieval request. In other words, this parameter plays the role of penalty to decrease the activation (the more similarity, the lower the penalty), whose magnitude is determined by $P$.

With the aforementioned similarity, the partial matching mechanism of chunk retrieval can be used. With regard to knowledge retrieval in ACT-R, a chunk that matches the retrieval request issued by the production module is retrieved from the declarative module (a request to retrieve chunks related to words containing ‘ri’ as a mora will result in chunks, e.g. ‘ringo ri head’). However, using partial matching, even chunks that do not precisely match the production module’s request will be retrieved if their activation (Equation (2)) is high at the time. This mechanism allows for flexible selection and the reproduction of specific errors.

In ACT-R, $M_i$ is defined by the modeler based on the assumption that it does not change with experience (i.e. the execution of the model). Thus, we treat this as an innate factor in language development based on the distinctive feature [13] for each combination of mora knowledge. In other words, we assume that the similarity between two morae can be defined by the degree of overlap of distinctive features derived from the organic structure. The calculation method is described in the next section.

### 3.2.2. Base level

The base level is the basic element of activation that corresponds to learning and forgetting. This model maps...
the base level to experiential factors in the development process (Equation (3)).

\[ B_i = \ln \left( \sum_{j=1}^{n} t_j^{-d} \right) + \beta_i \quad (3) \]

The value is computed from the number of presentations for chunk \( i \) (\( n \)) and the elapsed time as the chunk was referenced (\( t_j \)). \( d \) indicates the decay rate, and \( \beta_i \) is an offset constant that can be modulated according to the aim of the simulation. In other words, the more chunks used in the past, the higher the activation, and it decays according to the last time it was used. In the present study, we examine the role of such experiential factors in the development of phonological awareness. In particular, we assume that an increase in the base level corresponding to the accumulation of experience reduces the relative importance of innate factors (similarity).

3.3. Interactions for playing Shiritori

For a child with immature phonological awareness, it is difficult to play Shiritori following the standard rules. If the child answers with a word that does not conform to the Shiritori rules, such as ‘ri-n-go’ \( \rightarrow \) ‘o-ka-si’ (sweets), another error will usually occur if the error is not corrected. For children with immature phonological awareness to appropriately play Shiritori, the intervention of an adult is necessary [22]. In this study, when an error occurs during the task, the adult agent repeats the previous word (‘ri-n-go’, in the aforementioned case) to make the child try to determine appropriate mora connections from the word presented. Once a correct answer is given, the base-level learning of ACT-R is assumed to strengthen the appropriate word–mora relationship, eventually leading to matured phonological awareness.

4. Simulation

Using the aforementioned model, we performed four simulations. The first simulation was a replication of the previous study [28] to confirm the influence of knowledge similarity (innate factor) and base level (experiential factor) calculations on the generation of phonological errors. The second and third simulations demonstrated the process of generating phonological errors and factors that suppress such errors by sequentially executing the Shiritori task to further explore the experiential factor. The fourth simulation set different similarities of morae to examine the influence of an innate base on the development process.

4.1. Simulation 1: occurrence and suppression of phonological errors

4.1.1. Aims and settings of the simulation

Aims and parameters: To demonstrate how the model produces and suppresses errors in Japanese morae using mora knowledge similarity and base level calculations, we manipulated the similarity weight \( P \) (Equation (2)) and base-level offset constant \( B_i \) (Equation (3)) of the child agents. By varying these parameters, we can manipulate the balance of \( B_i \) (experiential factor) and \( P_i \) (innate factor) defined in the activation formula (Equation (1)). In the previous study [28], these two parameters were broadly manipulated. For demonstration purposes, the simulation shows results obtained through the combination of four levels of \( P \) (1, 10, 30, and 60) and three levels \( B_i \) (0.1, 10, and 20). For each combination of \( P \) and \( B_i \), we executed the model 100 trials, in which the two agents engaged a Shiritori task for 3,600 s in the ACT-R simulation time. Other ACT-R parameters set for the two agents are as follows: decay rate \( d \) = 0.5, and activation noises \( \varepsilon_i \) = 0.25. The chunks held by the child and adult agents are the same, although each combination of the chunks of the child is attached with similarity values. The rest of this subsection describes the preparation of these chunk settings.

Chunks in the model: The word knowledge in the model was constructed from words listed in the Japanese word database [30]. Based on the Shiritori rules, first, we selected 20, 544 nouns from the database, excluding homonym duplications and words consisting of only one mora, such as ‘ro’ (furnace) and ‘wa’ (ring). Next, we sampled 2, 054 words from the list to replicate the size of the child vocabulary. In this sampling, we picked one in ten words from the noun list that was sorted with the Japanese mora order.

For mora knowledge, 103 pieces of knowledge were defined based on Japanese morae. The combination of these morae was assigned with similarity computed from the distinctive features listed in Table 1.

Similarity computation: While there are several possible methods to compute similarities from such features, the computation in this simulation followed the previous simulation [28]. Table 3 lists five examples of computed similarities (cosine similarity) in this simulation, which are extracted from 10,506 combinations (103 \( \times \) 102) of morae. In this computation, values in distinctive features are converted from blank space, numerator, and denominator to 0, 1, and 2, respectively. In this way, this conversion sets blank features with the lowest priority because the pronunciation of blank features depends on the context and cannot be considered as the necessary condition of the phoneme. Thus, this computation is assumed to
represent the importance of the feature in defining the phoneme.

To represent each mora with the aforementioned numerically converted distinctive feature, Table 3 lists vectors of consonant and vowel phonemes that compose the mora. For vowel-only morae (e.g. ‘a’ and ‘i’), the consonant part of the vector was filled with averages computed for all the Japanese consonants. This filling manipulation follows the definition of mora (every mora has the same duration of pronunciation, in which consonant parts always precede vowel parts). Based on this definition, the absence of the consonant part is supposed to be filled with default values of the language.

Table 3 includes the column ‘IsYoon’ that specifies Japanese special mora known as yoon (contracted sound). The morae categorized with yoon have three phonemes (e.g. ‘sju’ and ‘zju’). To compute similarity between yoon and other morae, we set it as an independent binary feature instead of adding another set of distinctive features.

### 4.1.2. Results and discussion

To quantify errors in phonological awareness, we counted mora pairs that combined the ending mora of the answer of the adult agent and the initial mora of the answer of the child agent (e.g. caregiver’s answer: ri-n-go → child’s answer: go-ri-ra; go-go, the wrong answer ko-n-ro (stove) → o-ka-si: ro-o). Moreover, we calculated the entropy of the counted mora pairs to examine the degree of variation (convergence) of the generated mora pairs under each condition. The number of words answered during Shiritori was observed to check the performance of the Shiritori execution.

Figure 3 shows the type and number of mora pairs that appeared in each condition. Graphs that align vertically exhibit different values of \( P \), and graphs that align horizontally exhibit different values of \( \beta_i \). The horizontal axis of each graph is the list of mora pairs sorted based on the Japanese mora order (‘a–a,’ ‘a–i,’ ‘a–u,’ ‘i–i,’ ‘i–a,’ ‘n–n’), and the vertical axis represents the number of appearances of that mora pair. The correct number (e.g. ‘a–a’) is written in red and the incorrect number (e.g. ‘a–i’) in blue. The numbers shown in the graph denote entropy values, representing the variation in the mora pairs.

As shown in the figures, several incorrect pairs (blue) appeared when the value of \( P \) was small (\( P : 1, P : 10 \)). In other words, we confirmed that the incorporation of sound similarity leads to the misuse of mora knowledge. Moreover, as \( P \) increases, entropy decreases; i.e. the mora used converges to correct mora pairs as the value of \( P \).

| Rank | Mora | cons | syl | high | back | low | ant | cor | voice | cont | nas | student | dr | round | isYoon |
|------|------|------|-----|------|------|-----|-----|-----|------|------|-----|---------|---|-------|-------|
| 1    | sju  | 2    | 1   | 1    | 2    | 2   | 2   | 1   | 1    | 2    | 1   | 1       |   |       |       |
| 2    | zju  | 2    | 1   | 1    | 2    | 2   | 2   | 1   | 1    | 1    | 1   | 1       |   |       |       |
| 3    | su   | 2    | 2   | 2    | 2    | 2   | 1   | 1   | 1    | 1    | 1   | 1       |   |       |       |
| 4    | y99  | 2    | 2   | 2    | 2    | 1   | 1   | 1   | 1    | 1    | 1   | 1       |   |       |       |
| 5    | a    | 1    | 2   | 1    | 1    | 1   | 1   | 1   | 1    | 1    | 1   | 1       |   |       |       |
| 6    | a    | 1    | 2   | 1    | 1    | 1   | 1   | 1   | 1    | 1    | 1   | 1       |   |       |       |
| 7    | a    | 2    | 1    | 1    | 1    | 1   | 1   | 1   | 1    | 1    | 1   | 1       |   |       |       |
| 8    | a    | 1    | 2   | 1    | 1    | 1   | 1   | 1   | 1    | 1    | 1   | 1       |   |       |       |

Notes: The method of conversion follows the previous simulation [28]. cons = consonant; syl = syllable; son = sonorant; back = back; ant = anterior; cor = coronal; con = continuant; dr = delayed release.
increases. As shown in Equation (2), large $P$ degrades the activation of pairs with low similarity, making partial matches less likely to occur. Eventually, the mispronunciation of morae is suppressed by an increase in the $P$ value. Based on this mechanism, we suppose the role of $P$ as an individual difference in sensitivity to sound features. Pertaining to such an individual difference, ASD has symptoms related to abnormal sensation [11], and its diagnosis criteria admit ranges of symptoms across a continuous spectrum [31]. Thus, we consider that there is a possibility that the parameter $P$ represents the position of an individual on such a spectrum (the severity of such symptoms).

In addition, focusing on horizontal graphs (with different base-level offset constants $\beta_i$), we can see that entropies (the variance in the number of appearances of mora) decrease as $\beta_i$ increases. This trend is observed in conditions other than those with the lowest $P$ setting. In Equation (1), if $\beta_i$ is large, the effect of changes on other factors (e.g., similarity $P$) will be relatively small. In other words, the effect of the base-level (learning) on the suppression of phonological misuse is apparent.

Figure 4 shows the number of words correctly answered (the number of times Shiritori was continued) under each condition. Based on this graph, we can see that two parameters relate to the success of this task: suppressing innate similarity and accumulating lifelong learning increase the number of correct answers. Thus, the result suggests that innate and experiential factors are involved with the development of correct phonological awareness.

### 4.2. Simulation 2: process of learning through the Shiritori task

#### 4.2.1. Aims and settings of the simulation

This simulation was intended to demonstrate the increase in the base level caused by the actual task execution. In this simulation, a trial includes five Shiritori sessions of 600 s. Based on these sessions, the model updates $B_i$ according to Equation (3). Parameters manipulated in the previous section were fixed ($P$: 10, $\beta_i$: 0.1), and the other conditions were the same as in the previous section. We simulated 100 trials for each condition.

#### 4.2.2. Results and discussion

Figure 5 shows the entropy trend for each session as a function of the number of repeated sessions. Each series in this graph is the entropy of aggregating the mora pairs based on the first mora of the answer of the child agent. In other words, it indicates the variety of morae leading...
to a specific mora perception by the child (e.g. ri-n-go, ko-n-ro, or ne-n-do → o-ka-si). The series in the figure is categorized as vowels and different consonants with vowels. Based on the first session shown in the figure, we can observe a significant difference between these categories. The vowels have the highest entropy, indicating that the child agent connected the various end morae uttered by the adult agent to a vowel mora.

Although this trend is rapidly suppressed in the second session, this does not apply to learning on the task. As shown in Figure 6, the number of answers generated in the task decreases as the session progresses. In the later session, the child agent repeated the same word until the end of the session. This simple behavior can be considered as the cause of decrease in entropy shown in Figure 5. The low variety in the answers of the adult in the later session, caused by the discontinuation of the task, leads to the convergence of entropy.

One possible reason for such word repetition can be found in the previous study on ACT-R base-level learning. Lebiere [32] indicated that the default base-level learning (Equation (3)) leads to pathological repetitive behavior and extremely high activation in a specific chunk (a loop where one piece of knowledge becomes so active that it is constantly retrieved while ignoring others). In the current simulation, the erroneous experience (the experience of using confused mora) in the first session may lead to convergence to inappropriate specific chunks. The following simulation was intended to determine which type of chunks (words or mora) causes this erroneous process and what factors correct it.

4.3. Simulation 3: learning between sessions

4.3.1. Aims and settings of the simulation

To develop phonological awareness, the convergence to inappropriate chunks described in the previous section must be resolved. Thus, we included some interventions between session intervals. The assumption was that the intervention made by adults during the interval balances knowledge and prevents convergence to the incorrect use of knowledge. Even in a real-world environment, children do not learn phonological units only by playing Shiritori. They are guided by adults (a caregiver or a teacher) to engage in various types of language learning, e.g. writing and reading a table of the Japanese syllabary. This simulation aims to examine the effect of such learning on suppressing convergence on inappropriate chunks.

This simulation manipulated the activation of knowledge in the interval and its relation to success or failure in Shiritori. We incorporated a process to make the base-level activation of each chunk the same as the achieved maximum base-level activation value. We assume that this manipulation represents the idealized intervention that makes the child explore learning targets according to her/his level of understanding. To determine the cause of the failure presented in the previous simulation, this manipulation was applied to each type of knowledge (word, mora, and word–mora; Table 2). Other settings were the same as that of the simulation described in the previous section.

4.3.2. Results and discussion

Again, as in the previous section, we observed entropy in the mora pairs and the number of answers for each session. Figure 7(a–c) shows the entropy trend for each session. These are graphs for each type of knowledge that was manipulated. Figure 8 shows the number of correct answers per session.

As shown in Figure 5, the first session in Figure 7(a–c) exhibits a significant difference in entropy between categories, which rapidly converges to a specific mora. However, the trend in the number of answers (Figure 8) is different from the previous simulation, showing small decrease in the number of answers when we manipulated the activation of mora-related knowledge. In other words, the execution of Shiritori was improved, and the use of mora knowledge converged, suggesting that the intervention affects the correction of the inappropriate use of mora knowledge. However, Figure 8 indicates that
4.4. Simulation 4: effect of similarity setting

4.4.1. Aims and settings of the simulation

The previous two simulations explored how experiential factors affected the development of phonological awareness. To support the findings obtained so far, Simulation 4 examined the development of phonological awareness based on a different innate base. The settings were the same as those in Simulation 3, except the similarity setting adopted from one of the studies dealing with Japanese mora [14]. Table 4 lists examples of the similarity computation in this simulation. In contrast to Table 3, phonemes included in the mora (consonant, vowel, and special phonemes in the case of yōon) are not concatenated but averaged. Moreover, the numericalization of distinctive features is different. The distinctive features of blank space, −, and + are converted to 0, −1, and 1, respectively. Thus, this simulation treats blank features as a neutral priority and perceives several phonemes in a single mora as a single ‘chunk.’

4.4.2. Results and discussion

Figures 9(a–c) and 10 show the entropy trend of each session and the number of answers per session, respectively. The general trend of these results is consistent with Simulation 3 (Figures 7 and 8), showing a significant difference between categories in the first session and rapid suppression in the second session. However, we can determine the contrast of the mora type that exhibits the highest entropy between the two simulations. While the vowel morae (red lines) in Simulation 3 exhibited high entropy, having commonality with reported cases about children who experienced vowel confusion [11], the current simulation did not exhibit such tendency. This suggests that the innate similarity settings affect the confusion of morae in the initial stages of development.

5. General discussion and future work

With regard to research on SER, this study attempted to represent

(1) development process leading to erroneous phonological generation and
(2) experiential factors involved in the suppression of phonological errors.

Before exploring each problem, Simulation 1, replication of the previous study, was conducted to demonstrate how our model represents the development process with regard to the innate basis. Although the first simulation did not learn through experience, it shows the
Table 4. Examples of mora vectors and similarity using different calculation methods [14].

| Rank | Mora | Vocalic | Consonantal | High | Low | Back | Anterior | Coronal | Round | Tense | Voice | Continuant | Nasal | Strident |
|------|------|---------|-------------|------|-----|------|----------|---------|-------|-------|-------|----------|-------|----------|
| 1    | ha   | 0       | -1          | -1   | 1   | -1   | -1       | -0.5    | 0.5   | -0.5  | 0.5   | -0.5     | -0.5  | -0.014   |
|      | hja  | -0.333  | -1          | -0.333 | -0.333 | 0.333 | -1       | -0.667  | 0     | 0.333 | -0.333 | -0.333   | -0.333 | -0.014   |
| 31   | ki   | 0       | 1           | 0    | 0   | -1   | -1       | -1       | -0.5  | -0.5  | -0.5  | -0.5     | -0.5  | -0.04    |
|      | gi   | 0       | 1           | 0    | 0   | -1   | -1       | -1       | -0.5  | -0.5  | -0.5  | -0.5     | -0.5  | -0.014   |
| 623  | do   | 0       | 0           | 0    | -1  | 0    | 0        | 0        | 0.5   | -0.5  | -0.5  | -0.5     | -0.5  | -0.04    |
| 752  | a    | 1       | -1          | 1    | 1   | -1   | -1       | -1       | -0.5  | -0.5  | -0.5  | -0.5     | -0.5  | -0.158   |
|      | ka   | 0       | 0           | 1    | 0   | -1   | 0        | 0        | -0.5  | -0.5  | -0.5  | -0.5     | -0.5  | -0.172   |
| 3587 | a    | 1       | -1          | 1    | 1   | -1   | -1       | -1       | -0.5  | -0.5  | -0.5  | -0.5     | -0.5  | -0.444   |
|      | i    | 1       | -1          | 1    | -1  | -1   | -1       | -1       | -0.5  | -0.5  | -0.5  | -0.5     | -0.5  | -0.444   |

Figure 9. Changes in entropy using mora knowledge in Simulation 4; results based on different similarity settings. (a) word knowledge manipulation (b) mora knowledge manipulation (c) word–mora knowledge manipulation

Figure 10. Changes in the number of answers in Simulation 4; results based on different similarity settings (error bar is ±SD).

The cases of erroneous pronunciation observed in children with ASD in the real world. In related studies [9, 15, 16], it has been reported that children with ASD face difficulty in becoming aware of consonants. For example, Grandin who is a behavior scientist with ASD reported her personal experience as follows: ‘To me, cat, hat, and pat sounded the same, because those consonants are quick.’

The process shown in Simulation 2 represents that such an innate basis of a child with ASD leads to erroneous confusion of mora, and then, the memory activation mechanism strengthens such errors. Moreover, repetitive behavior observed in the simulation is included in the criteria of ASD diagnosis [31]. These commonalities suggest that the current ACT-R model can help understand a specific pattern of the human symbol emergence process in the real world and eventually support activities to improve such erroneous behaviors.

Based on the findings, the third simulation tried to address the second objective of this study to determine factors that can correct erroneous processes. By directly manipulating the interval learning, we found that balancing the use of mora was important to recover and acquire proper phonological awareness. We consider that such manipulation corresponds to training for kana literacy, conducted in the first grade of Japanese elementary schools [33]. The activity of Japanese class in that grade includes practicing pronunciation of each kana character and finding words that include each kana character. Thus, our simulation suggests that if such training...
for kana literacy comprehensively activates the overall mora knowledge, errors caused by innate bases can be resolved. However, in this simulation, learning in the interval was conducted by direct manipulation of the parameters. In the future, it will be necessary to comprehensively study and implement interval learning for the simulation of real-world educational settings. By constructing a setting that corresponds to the real world, we can explore detailed conditions of comprehensive activation of phonological knowledge. In addition, we need to evaluate such simulations to confirm that the proposed model is a suitable representation of the real phonological awareness formation process. This can be done by comparing the model with data pertaining to the classification of phonological errors in development [34].

To summarize, our simulations successfully represented an erroneous developmental process (as repetitive behavior started from innate bases) and a factor of suppressing such process (as balancing knowledge activation). The important assumption of our model is that the activation mechanism of chunks governs the development of phonological awareness. We consider that these representations based on a general cognitive theory can contribute to understanding various symbol emergent processes in real humans and eventually leading to help resolve development disorders.

There are other unanswered problems about this study. We assume that language development emerged from the innate base, which is represented as the distinctive feature. However, as presented in the settings of Simulations 1 and 4, the influence of the innate base on phonological awareness (activation of mora chunks) has changed with the method of similarity computations. As discussed above, the similarity computed in Simulation 1 leads to more erroneous pronunciation reported in ASD (consonant deletion). Based on the result, though further research is required to specify it, we can speculate that some of the assumptions drawn in Simulation 1 reflect genetic traits of ASD. In the future, we will need to verify the assumptions of innateness of this setting.

Finally, we remark on the contribution of this simulation to the SER approach. In our simulation, we implemented symbols (words and morae) in the system as declarative knowledge and did not deal with the real environment. Instead, we used knowledge of vocal sounds (distinctive features), which were developed in linguistic fields. This approach does not conform to the paradigm of the conventional SER technique, which focuses on self-organizing symbol emergence processes. With regard to the SER context, our simulation contributes to the understanding of how the already emerged symbols were generated from the innate base using a dynamic process. Symbols used in the simulation are socially defined and can be easily understood by community members. We believe knowledge obtained from such simulation contributes to the future development of the SER technique, which is an endeavor conducted by human researchers, by providing a blueprint of the emerging system.

Notes

1. The present study follows this stance of original universal phonetics because of simplicity of the explanation. However, it is needed to acknowledge that this stance has been challenged by several researchers [15, 16]. These researchers focusing on evolutionally and emergent aspects of language development while not completely denying universality of basic phonetics.

2. The same error of deleting consonants is reported in studies with other languages. An English report by Grandin [11] discusses how children with autistic spectrum disorder (ASD) may ignore consonants because they have difficulty in becoming aware of voiceless sounds.

3. This type of formulation (a linear combination of factors) can manipulate factors that independently affect the human memory process.

4. In this manipulation, we assume that the offset value represents the average amount of lifetime learning for the chunks of the agent. An overall effect of learning is supposed to increase as the individual grows up.

5. ACT-R estimates how long humans take to accomplish a specific task by allocating primitive time to each component of an activity (firing production rule, retrieving chunks, processing aural stimuli, and verbalizing syllables). In simulation studies using ACT-R, it is common to use this setting the length of task trial.

6. It is estimated that a four- to five-year-old child has a vocabulary of 1, 500—2, 400 words[21].

7. For standardization purposes, we use cosine similarity of two vectors.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by JSPS KAKENHI [grant number JP18H05068] and [grant number JP20H04996].

Notes on contributors

Jumpei Nishikawa received Bachelor of Informatics degrees from Shizuoka University in 2020. He is pursuing master’s degree in Graduate School of Integrated Science and Technology, Shizuoka University. He is interested in the cognitive mechanism of language development and its modeling.

Junya Morita received the Ph.D. degree from Nagoya University in 2006. He served as an assistant professor of the School of Knowledge Science, Japan Advanced Science and Technology (2006–2014) and as a designated assistant professor of the Institute of Innovation for Future Society, Nagoya University (2014–2016). He is now an associate professor of the
School of Informatics, Shizuoka University. His research aims to develop a novel integration between humans and machines based on technologies from computational cognitive modeling and affective computing.

References

[1] Taniguchi T, Nagai T, Nakamura T, et al. Symbol emergence in robotics: a survey. Adv Robot. 2016;30 (11-12):706–728.
[2] Pinker S. The language instinct. New York (NY): William Morrow & Co; 1994.
[3] Skinner BF. Verbal behavior. New York (NY): Appleton-Century-Crofts; 1957.
[4] Baron-Cohen S. Mindblindness: an essay on autism and theory of mind. New York (NY): Oxford University Press; 2007.
[5] Tomasello M. The cultural origins of human cognition. Cambridge (MA): Harvard University Press; 1999.
[6] Mielke J. The emergence of distinctive features. New York (NY): Oxford University Press; 2008.
[7] Shain C, Elsner M. Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers). Minneapolis (MN): Association for Computational Linguistics. 2019. p. 69–85.
[8] Trubetzky NS. Principles of phonology. Oakland (CA): University of California Press; 1969.
[9] Anthony JL, Francis DJ. Development of phonological awareness. Curr Dir Psychol Sci. 2005;14(5):255–259.
[10] Kobayashi H. Oninshiki no keisei to kotoba no hattatsu – ‘kotoba ga osoi’ wo kangaeru – [Phonological awareness formation and language development – considering ‘language delay’ –]. Nagano: Kodama Shuppan; 2018.
[11] Oishi N. Hyoka [Evaluation]. In: Ishida H, Ishizaka I, editors. Gengochakakushi no onimpuresu no henka [Phonological process and short-term inhibition. In: Proceedings of the 31st Annual Virtual Meeting of the Cognitive Science Society 2020. 2020. p. 571–577.
[12] Taatgen NA, Anderson JR. Why do children learn to say ‘broke’? A model of learning the past tense without feedback. Cognition. 2002;86(2):123–155.
[13] Van Rijn J, Van Rijn H, Hendriks P. Cognitive architectures and language acquisition: a case study in pronoun comprehension. J Child Lang. 2010;37(3):731–766.
[14] Matusevych Y, Schatz T, Kamper H, et al. Evaluating computational models of infant phonetic learning across languages. In: Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society 2020. 2020. p. 571–577.
[15] Bothell D. ACT-R 7.21+ reference manual. 2020.
[16] Amano S, Kobayashi T. Kihongo database: gogibetsu tango/sumitsudo [Basic word database: word familiarity by meaning]. 2009. Japanese.
[17] Malt BZ, Zec F. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Arlington (VA). 2013.
[18] Leibler C, Best BJ. Balancing long-term reinforcement and short-term inhibition. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society. 2009. p. 2378–2383.
[19] Ministry of Education, Culture, Sports, Science and Technology. Shogakko gakushu shido yoryo [Study of course content] 2013. Japanese.
[20] Nakamura T, Kojima C, Fujiyama Y. Kenjohattatsu ni okeru onimpurosesu no henka [Phonological process decline in two- to six-year-olds]. J Rehabilitation Sci. 2015;10:1–13. Japanese.