CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

The acquisition of the semantics of Japanese numeral classifiers: The methodological value of nonsense

Maki KUBOTA[1] [ID], Yuko MATSUOKA[3] and Jason ROTHMAN[1,2]

[1]AcqVA Aurora Center, UiT the Arctic University of Norway
[2]Centro de Investigación Nebrija en Cognición, University of Nebrija
[3]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh
Corresponding author: Maki Kubota; Email: makikubota5@gmail.com

Abstract

This study examined the acquisition of numeral classifiers in 120 monolingual Japanese children. Previous research has argued that the complex semantic system underlying classifiers is late acquired. Thus, we set out to determine the age at which Japanese children are able to extend the semantic properties of classifiers to novel items/situations. Participants completed a comprehension task with a mouse-tracking extension and a production task with nonce and familiar items. While the comprehension results showed ceiling effects on familiar and nonce items, age significantly modulated a difference in accuracy between familiar and nonce items in the production task. The findings suggest that the acquisition of the underlying semantic system is acquired much earlier than previously argued. Previously attested issues with Japanese classifier production in young(er) children are more likely to reflect accessing difficulties than indexing the underlying grammatical competence of the classifier system.

1. Introduction

Children begin to use language to categorize and count items in their environment early in their development. To do so, most languages only require combining nouns with numerals to construct numeral noun phrases (e.g., 'three lions'). However, some languages such as Japanese, Chinese, and Korean have special expressions accompanying the numerals to categorize items that they quantify (e.g., Japanese: san *too* no raion 'three *CL*-GEN lion'). These are called NUMERAL CLASSIFIERS. Numeral classifiers behave similarly to collective nouns, which comprise an open class of words that quantifies mass unit (e.g., herd of cattle, flock of birds). Although much work has examined children's acquisition of classifiers, there is little consensus as to HOW and WHEN children acquire their syntactic patterns and the corresponding semantic system. In order to ensure that children can assign meanings to a classifier (e.g., *too* categorizes large animals) – rather than merely

associating specific items to particular classifiers (e.g., *too* being lexically associated with the noun *lion*) – it is crucial to test whether they can extend the use of classifiers to novice items/contexts. The current study aims to examine this question by using familiar and nonce items in both comprehension and production modalities to map the developmental trajectories of classifier acquisition in Japanese monolingual children.

Count syntax serves as a cue for individuation in English and other Indo-European languages. It involves using words directly with numerals (e.g., one cat) or in singular or plural forms (e.g., a cat, cats), as well as quasi-cardinal determiners (e.g., these/those cats). These forms indicate reference to sets of countable things. On the other hand, languages like Chinese and Japanese do not employ count syntax. Instead, they utilize classifiers to explicitly indicate reference to individual entities – (Li et al., 2008). In Japanese, numeral classifiers must be attached to a noun whenever the quantity is specified, but classifiers are not obligatory on every noun/adjective unlike grammatical gender or number (Aikhenvald, 2000). Although there are approximately 150 Japanese numeral classifiers, only about 30 are found in frequent, daily use (Downing, 1996). The semantic properties of each classifier category and its organization is complex, opaque, and differs across various classifier languages (although a certain universality has been attested such as animacy and shape; Adams & Faires Conklin, 1973). In Japanese, classifiers are strictly divided into two major categories: animate and inanimate. Japanese classifiers (unlike Chinese) do not allow animate and inanimate items to belong to the same classifier. They are further organized around semantic features, such as type of animal, shape, and function as in Figure 1 below (Yamamoto & Keil, 2000, p.381). For example, *–hon* is the Japanese classifier for long, thin items and *–mai* is used for flat, thin items, while *–ri is* used to count humans and *–hiki* for small animals and insects. Numeral classifiers tend to be associated with the qualities or properties of referents, or with the ways in which we relate to those referents, rather than directly with the nouns that refer to them (Jarkey & Komatsu, 2019). Some nouns, however, are strongly associated with a particluar classifier, and so using a different classifier is typically judged as unacceptable. In Japanese, 'general classifiers', such as *–tsu* or *–ko,* can be applied to a wide range of inanimate nouns that vary across dimensions (however, there are several nouns in which the use of general classifiers is not appropriate e.g., using -ko for trains).

Children can quickly map novel nouns to meaning/concepts long before they begin to acquire the classifier system; albeit both involve classification of entities (Uchida & Imai, 1999). Although considerable work has refined our understanding regarding the age and conditions under which acquisition of classifiers take place, not all evidence points in the same direction – while some studies show that common classifiers are generally fully mastered by the age of six (Sanches, 1977; Sumiya & Colunga, 2006; Uchida & Imai, 1999), others claim that even older children well beyond schooling age do not show full acquisition of basic classifiers (Matsumoto, 1987; Salehuddin & Winskel, 2009). Such inconsistencies can be attributed to several factors, the most obvious one being the differences in testing modality between comprehension and production abilities. To begin with, there are very few studies that examine comprehension and production of classifiers in child development. The most recent study (Hao et al., 2021) found a significant gap between comprehension and production accuracy of classifiers in Mandarin-speaking children, corroborating the findings of previous studies (Chien et al., 2003; Uchida & Imai, 1999) that comprehension precedes production in classifier acquisition. Among those studies that examine comprehension of classifiers, children seem to reach ceiling (over 90% accuracy) on various types of sortal classifiers by age six, with significant improvements between four to five years of age (Japanese: Uchida & Imai, 1996, 1999; Sumiya & Colunga, 2006; Yamamoto & Keil, 2000, Chinese: Chien at al., 2003;
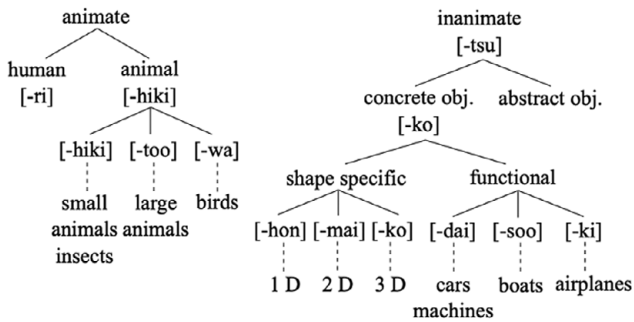
Figure 1. Japanese numeral classifier system (taken from Yamamoto & Keil, 2000, p.381).

Li et al., 2010; Hao et al., 2021). On the other hand, those that test the production of classifiers show strikingly low accuracy for six-year-old Chinese and Japanese children (Hao et al., 2021; Matsumoto, 1987; Uchida & Imai, 1999) and a study with Malay children (Salehuddin & Winskel, 2009) shows that even children as old as nine could only produce the correct classifier around 50 percent of the time. Of course, it cannot be assumed that all Asian numeral classifier systems should evidence the same acquisition patterns, not least because specific languages are more and less transparent relative to one another (e.g., as highlighted above, Chinese is less transparent than Japanese in that only the former permits a single classifier to collocate with both animate and inanimate nouns). Yet, it is worth pointing out the general trend nonetheless: studies that rely on production data regardless of the specific language suggest more pronounced protracted development.

Despite differences in the developmental trajectories between comprehension and production of classifiers, the overall acquisition pattern suggests that children first recognize the grammatical function of classifiers by learning the appropriate classifier for each noun via rote association from input. They may, therefore, simply insert (in theory) any classifier that they have come across in their input without extracting any semantic rules. At some point in their development, they begin to extract semantic rules with enough exposure to examples from each classifier category, and eventually learn to apply this rule to entities that share similar semantic properties (Uchida & Imai, 1999). Thus, the acquisition of classifiers involves the process of pairing classifiers to nouns as well as coming to intuit and form rules based on the underlying semantics. Only after the latter, can one expect children to apply classifier forms to novel contexts/entities in a target manner. This has been argued to be a protracted learning process: "learning the full semantic system still takes a long time" (Uchida & Imai, 1999, p.66). Most previous studies (as cited above), however, use real items that are available in the child's input to test their use and knowledge of classifiers. Although some attempts have been made to uncover the relationship between semantic categorization skills and classifier knowledge by running correlational analyses (Hao et al., 2021; Sera et al., 2013), it is difficult to tease apart the grammatical vs. semantic knowledge of classifiers via a paradigm that tests only known items. An exception is a study by Li et al. (2010) which tested Mandarin children's comprehension of numeral classifiers by using both familiar and novel items. The results showed no differences in accuracy between familiar and novel items even for children as young as age three (although their accuracy was 56% overall with 33% as chance level), suggesting that children are able to extract semantic information from classifiers at a relatively young age. Another exception comes from a training study by Uchida and Imai

(1999) which tested four- and five-year old Japanese children's ability to learn the semantic rule of a classifier -*too* (big animals) by undergoing two different types of training: (a) explicit instruction; or (b) exemplar-only explanation. The findings demonstrated that while four-year olds did not make use of the examples provided by the researcher to generalize the meaning of -*too* to novel items, five-year-olds were successfully able to do so, performing on par with their peers who received explicit instruction.

Given that (a) Li et al. (2010) and Uchida and Imai (1999) only examined the COMPREHENSION of classifiers in novel contexts and (b) the PRODUCTION of classifiers documented in a larger literature attests a more protracted trajectory than comprehension, it is crucial to test children's knowledge of classifiers in familiar and novel contexts in BOTH modalities. In doing so, we can better gauge how and when semantic knowledge of classifiers is acquired in childhood, while shedding light on if (and why) the modality of testing – comprehension versus production – matters for attaining the evidence needed to answer this query. In the current study, we investigated Japanese monolingual children's processing, knowledge, and use of classifiers in novel and familiar contexts by employing a mouse-tracking/comprehension and a production task. We incorporated a mouse-tracking extension in the comprehension task to examine whether there are any discrepancies between behavioral accuracy and real-time processing of classifiers. Combining online and offline measures and showing replication across modalities or lack thereof, will contribute further to revealing how modality affects acquisition patterns of classifiers.

Following this, we formulated the following research questions:

1. What is the developmental trajectory of comprehension and production of classifiers in Japanese monolingual children? What is the critical age in which the use and knowledge of classifiers is acquired?

We hypothesize that children's use and knowledge of classifiers will undergo rapid development around the age in which they enter the formal schooling (from age six to seven), given that they will receive more quantitatively and qualitatively rich input from the environment.

2. Does familiarity (nonce, familiar) and animacy (animate, inanimate) modulate the development of classifier production and comprehension?

We hypothesize that both familiarity and animacy will modulate the development of classifier production and comprehension (as stated in detail in Section 2.5 Statistical Analysis, there should be a significant interaction between Age and Familiarity as well as Age and Animacy). Namely, we expect children to perform better on familiar than nonce as well as animate than inanimate items until around middle childhood (age ten to twelve) in which they should perform at ceiling (i.e., fully acquire the target classifier system) regardless of familiarity or animacy. There may also be a three-way interaction between Age, Familiarity, and Animacy, in which the developmental differences between familiar and nonce items may be present only in the animate or inanimate items.

2. Methods

2.1. Participants

145 children participated in this study via an internet-based experiment platform (Gorilla). Six participants' data were excluded due to poor data quality. This study was approved by Norwegian Center for Research Data (Ref number: 309414). Five other participants' data

Table 1. The number of participants in each group

Age group	Number
3	17
4	18
5	19
6	37
7	12
8	4
9	3
10	3
11	7

were removed because they indicated language impairment or developmental disorders. An additional 14 participants' data were excluded due to them not fitting the age criteria, leaving the final pool of participants as 120 children (Mean age = 6.21, SD = 2.05, Range = 3.05 – 11.83, Female = 62). The numbers of participants in each group are presented in Table 1. They were all native monolingual speakers of Japanese living in Japan with Japanese parents. Their Socio-Economic Status (SES) was measured via the mother's final education from a scale of 1 to 5 (Mean = 4.52, SD = 0.71, Range = 3 – 5). The participants were recruited through a Japanese online recruitment platform (Lancers).

2.2. Materials

We tested Japanese children's knowledge and use of classifiers by administering a comprehension task with a mouse-tracking extension and a production task, both built and run in Gorilla. For both tasks, we tested the children on six classifier categories: *-hon* (long, thin), *-mai* (flat, thin), *-dai* (machine), *-ri* (humans), *-hiki* (small animals), *-wa* (birds). These classifiers were chosen since they are frequently used in modern Japanese, have well-defined and salient perceptual features, and are visually distinct and common in ordinary life.

For the comprehension task, there were 48 target items in total, with 24 familiar items and 24 nonce items. There were 4 items in each classifier category for familiar and nonce items, as presented in Table 2. The frequency of the familiar items was matched within each classifier category, but not across categories. Indeed, it was extremely difficult to match the frequency across categories, due to the fact that some categories (such as *-dai* for machines and *-wa* for birds) only occur with specific items/animals that are not as frequent in the input compared to those that belong to more general categories such as *-hon* (long, thin) or *-hiki* (small animals).

The nonce labels for the nonce items were normed with 46 adult native speakers of Japanese. They were asked to rate how much meaning the words presented to them carried, on a scale of 1-4 (1 = there is no meaning to this word, 4 = the meaning of this word is clear). All nonce labels used in Table 2 scored below 1.4. The pictures used for familiar and nonce items were normed twice with 49 adult native speakers of Japanese in the first round and 59 adult native speakers of Japanese in the second round. In both

Table 2. Full list of classifier items in the comprehension task
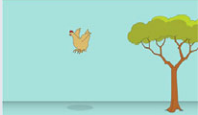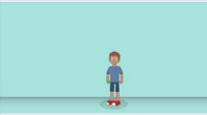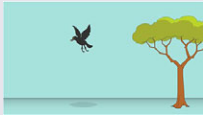
Familiar items		
Inanimate Classifiers		
Hon (1D long thin)	**Mai (2D flat)**	**Dai (machines)**
banana#	leaf#	TV#
pencil#	map#	phone#
carrot	towel	bicycle
rope	plate	car#
Animate Classifiers		
Ri (humans)	**Hiki (animals)**	**Wa (birds)**
girl#	cat#	chicken#
boy#	dog#	crow#

Table 2. *(Continued)*
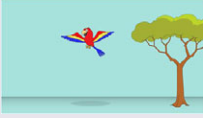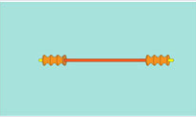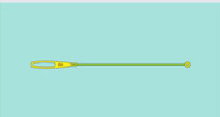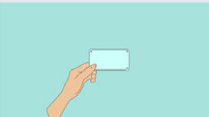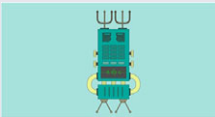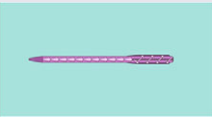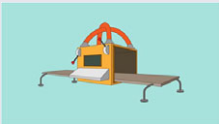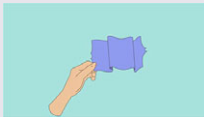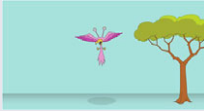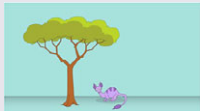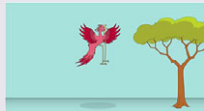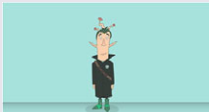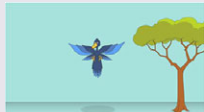
Familiar items		
man	mouse	crane
woman	fish	parrot

Nonce items		

Inanimate Classifiers

Hon (1D long thin)	**Mai (2D flat)**	**Dai (machine)**
sonu#	poru#	naso#
yapu#	mupi#	koni#
honi	nopu	gemi
chiza	napu	nefu

Table 2. *(Continued)*

Familiar items		
Animate Classifiers		
Ri (humans)	**Hiki (animals)**	**Wa (birds)**
memu#	rido#	sako#
yopo#	romo#	fuma#
yupi	tasa	reni
ropu	suro	tapo

#= subset of items used in the production task

rounds of piloting, we asked them to write the appropriate classifier for each picture. In the first round, there were some nonce pictures that had low classifier agreement (ranging from 8% to 100% agreement) while the familiar pictures all elicited 80% agreement or higher. Thus, we discarded those that elicited less than 70% agreement and created new nonce stimuli. These new stimuli/pictures (along with those that were kept in the first round) were tested again with a new set of raters. Pictures that reached higher than 70% agreement in the second round were used in the experiment (as presented in Table 2).

2.2.1 Production task
In the production task, children were instructed to count the number of items depicted in a picture. The number of items always followed the sequence of one, two, and three. First, the participants watched a video of the researcher explaining the task in Japanese and underwent a microphone check to make sure that the recording system worked. Then, the
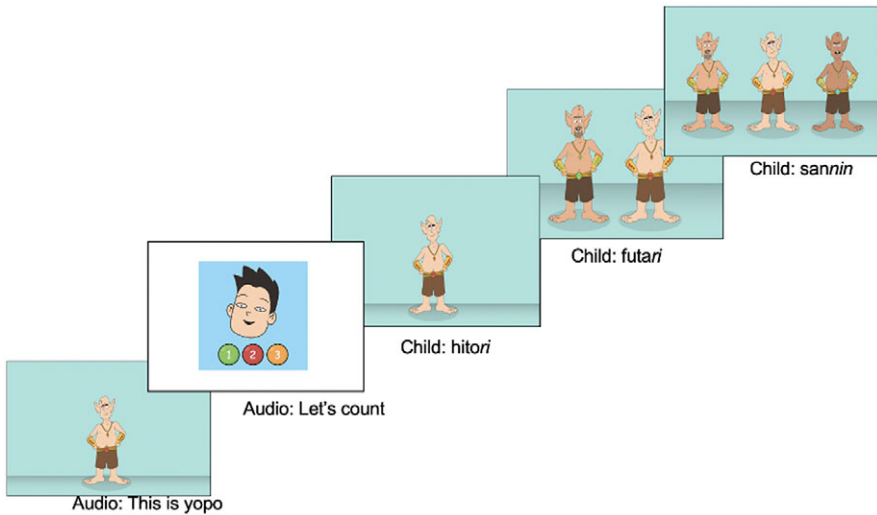
Figure 2. Illustration of the production task.

participants did two practice trials using classifiers that differed from the target classifiers (i.e., *-too* for large animals and *-hai* for a glass/cup of liquid). In each trial, they were presented with the name of the item (e.g., "kore wa yopo desu これはヨポです" *This is a yopo*). Then they were asked to count the items presented one by one on the screen as depicted in Figure 2. After counting the items, they were asked to describe two sets of pictures in which we also tested their knowledge of passive structures as a filler of sorts (note that passives are formed in Japanese with a dedicated, overt morphological exponent) and with the goal of using it for an additional study (thus the passive data are not included as part of the current study).

In the production task, we used half of the items in the comprehension task (N = 24, 12 familiar items and 12 nonce items) which are presented in Table 2 with an asterisk. This measure was taken to shorten the length of the experiment and minimize fatigue effects. The trials were randomized across participants. The production task was programmed in Gorilla such that the recording automatically started when the picture was presented to the participants and ended when the participant clicked on the button to move on to the next picture/trial. There was a break after half of the trials were completed.

Four research assistants who are native speakers of Japanese and have training in linguistics transcribed and coded the data. One of the principle investigators (PI), who is also a native Japanese speaker, did a final quality control check – that is, went through all the transcriptions and their coding to check for any inconsistencies. Any inconsistencies or disagreements were resolved among the PI and the research assistants.

We first coded the data with a binary choice (0,1) depending on whether the child produced the target classifier (or not). Phonological errors such as saying "*sanpon*" instead of *sanbon* were coded as a target response (i.e., 1) as long as the child produced the target classifier. As discussed in greater detail below in the procedures section, given the age of the participants, a parent was required to supervise for those under 12. Although the parents were instructed to not interfere with the self-contained automated experiments, if they did, we would know since the sessions were recorded. All utterances

in which a parent (always a mother in our sample) provided the classifier to the child – for example, by interjecting and asking the child: "How many CL?" – were removed from the analysis. While occurring, parental interjection was quite rare, consisting of 0.6% of the data. A further 2.5% of the data that were unintelligible or contained no audio were also excluded from the analysis. All utterances coded as non-target were further categorized into three non-target response types, one denoting omission and two denoting sub-classes of commission: (a) NC = No classifier; (b) GC = general classifier; (c) WC = wrong classifier. NC indicates that the participant did not use a classifier and produced only numerals (e.g., ni instead of ni-*hon*). Responses are coded as GC when participants used one of the general classifiers: -*ko* or -*tsu* (e.g., ni-*ko* instead of ni-*hon*). WC were coded when the participant produced a non-target classifier (e.g., ni-*mai* instead of ni-*hon*). We should note here that some of the responses in GC and WC are not necessarily an "error" or "ungrammatical" – for instance, it is generally accepted to use the general classifier -*ko* and -*tsu* for novice objects that are small and tangible (in the case of nonce items), or to use -*hiki* instead of -*wa* for certain birds.

2.2.2 Comprehension task

In the comprehension task, the participants were instructed to choose the appropriate picture that corresponds to the target classifier out of two options. Since the comprehension task included a mouse-tracking extension, the participants had to maximize their browser to full-screen mode to proceed to the comprehension task. Each trial consisted of the participant clicking on the small alien at the bottom of the screen to start the audio (i.e., numeral + CL) as they moved the cursor to click on either the top left or the top right picture. For instance in Figure 3, the participants hear "ichi-*mai* (one flat-CL)" after clicking on the alien and they have to move their cursor to the right to click on the corresponding target picture (i,e., plate). No time limit was set for each trial – the participants automatically moved onto the next trial when they clicked on a picture. They were instructed to click on the picture as quickly and accurately as possible.

As indicated in Table 2, there were 80 items in total, with 48 target items (24 familiar and 24 nonce) and 32 fillers. The full list of stimuli can be found in the Supplementary Materials. The target and the competitor were always either exclusively familiar or nonce items. We did not include familiar-target & nonce-competitor (or vice-versa) pairs, since children were likely to bias towards the familiar item when presented together with a nonce item, regardless of what the classifier was. We also counterbalanced the animacy of the target-competitor pairs. One-fourth of all trials were: (a) target-animate & competitor-inanimate (mismatched pairs); (b) target-inanimate & competitor-animate (mismatched pairs); (c) target-animate & competitor-animate (matched pairs); (d) target-inanimate & competitor-inanimate (matched pairs) respectively. We manipulated this factor given that Yamamoto and Keil (2000) found Japanese children to perform better when the animacy of the target-competitor pairs was not matched[1]. The position of the target picture/item (right or left upper screen) was counterbalanced, and the trials were randomized across participants. The comprehension task began with an explanation video of the task followed by two practice trials, in which no target classifiers were included. There was a break in between half of the trials. In the comprehension task, we

---

[1]As per suggested by the reviewer, we included the accuracy and reaction time of the target-competitor pairs (matched or mismatched) in the Supplementary Materials.
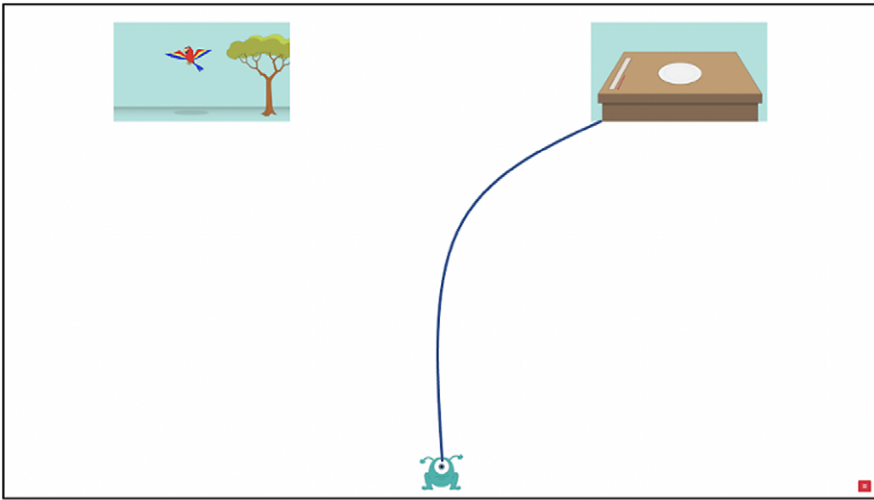
Figure 3. Illustration of the comprehension task.

analyzed their accuracy (i.e., whether they chose the target picture or not), and reaction time, as well as two common mouse-tracking measures: maximum absolute deviation (MAD) and sample entropy. MAD is computed as the largest perpendicular deviation between the actual and the idealized trajectory between starting and response positions. MAD assesses the degree of attraction toward an unselected response magnitude of activation for each response option as the decision process unfolds over time (Hehman et al., 2015). Sample entropy quantifies the degree of unpredictability of movement along the x-axis and is a measure of spatial disorder and complexity of the movement.

2.3. Mouse-tracking pre-processing

The mouse-tracking output from Gorilla provides a data file containing the coordinates of the participant's mouse position on the screen over time, with time stamps for each position. All coordinates are measured in pixels from the bottom-left edge of the screen (0,0), with the x-coordinate increasing as the mouse moves right and the y-coordinate increasing as the mouse moves up. Since this experiment was conducted online and participants have different devices and screen sizes, we used the normalized x-coordinate and y-coordinate values which consider the proportion of the screen space, making these coordinates comparable across different participants. We used the mousetrap package (Kieslich et al., 2019) in R to analyse the mouse-tracking data. We first filtered the data so that it only includes accurate trials, and only kept trials where mouse positions varied and also removed duplicated time stamps (1.1% data were removed). Trajectories were remapped so that all right-ending trajectories were remapped to left. Cursor's starting point was aligned by shifting the trajectories. Since the sampling rate differed depending on the participant's device, we interpolated trajectories so that each is represented by the same number of positions (101 steps) separated by constant time interval (i.e., time-normalization). The dependent variables (MAD and sample entropy) were calculated based on raw trajectory measures.

2.4. Procedure

As previously mentioned, the experiment was conducted remotely via Gorilla and the participants took part in the study in their own homes. We made sure that they could only access the experiment from their laptop or their computer (and not their phones or iPads). The participants first watched a general introduction video which instructed them to be in a quiet environment with no distractions. The parents were instructed to not provide their children with answers, and we also asked them to supervise their children if they were under 12 years old. After the parents signed the consent form, the children watched a short animation cover story video that involved two astronauts, Ken and Lisa, who get their spaceship stolen by aliens. The children were asked to help Ken and Lisa take back their spaceship by completing multiple missions that involve different creatures. We always administered the production task before the comprehension task, since the comprehension task gives away the target classifier and we wanted to avoid any learning effects that arise from this to influence the children's performance in the production task. Upon completing the production and comprehension tasks, parents filled out a language background questionnaire and a compensation form. The participants were compensated with a 1000-yen gift card via email. The entire online experiment can be accessed through the Gorilla open materials page https://app.gorilla.sc/openmaterials/686845.

2.5. Statistical analysis

We ran a generalized linear mixed effects model for binary dependent measures (accuracy on comprehension and production tasks), poisson generalized linear mixed effects model for count measures (non-target responses type) and a linear mixed effects model for continuous numerical dependent measures (reaction time on comprehension task and MAD and sample entropy values on mouse-tracking) using the lmer package (Bates et al., 2015) in R (R Core Team, 2021). We included age, familiarity (nonce or familiar), and animacy (animate, inanimate) as well as the interaction between age and familiarity and a three-way interaction between age, familiarity, and animacy into each model. Participants and items were included as random intercepts and familiarity as a by-participant slope. Age was centered around the mean. In order to investigate the critical age in which the use and knowledge of classifiers is acquired, we also used a modeling technique called Conditional Inference Trees (CTrees) (Breiman, 2001) using the partykit function (Hothorn et al., 2015) in R. CTrees first tests the significance of each independent variable, then the variable with the strongest association with the response is chosen, and a binary split is performed on the independent variable, which divides the dataset. The process then repeats on the subsets of the data and tests the remaining independent variables. We included age and familiarity as predictors since we were most interested in the interaction between these two variables. The outcome of the test can be visualized graphically as a tree, with the most important independent variable located at the top of the tree with further associations between independent variables shown lower down. P values and significant results are obtained via permutation, a resampling process similar to bootstrapping (see Levshina, 2015, p.291 for further details).

3. Results

We will first describe the results of the production task, then the behavioral results of the comprehension task, and lastly the mouse-tracking results.

3.1. Production task

3.1.1 Descriptive analysis

The accuracy of the production of classifier types split by animacy and familiarity is presented in Figure 4. Overall, it appears to be the case that children perform better on familiar than nonce items (familiar: $M = .31$, $SD = .46$, nonce: $M = .26$, $SD = .44$), and also perform better on animate than inanimate items (animate: $M = .34$, $SD = .47$, inanimate: $M = .22$, $SD = .42$). The accuracy of classifier production split by each age group, animacy, and familiarity can be found in the Supplementary Materials.

3.1.2 Generalized linear mixed effects (glmer) model

The summary output of the glmer is summarized in Table 3. Here, we are interested in examining the interaction between the predictors, specifically between age and familiarity, as this two-way interaction informs us about whether the development trajectories of familiar and nonce items differ and to what extent. We see a significant two-way interaction between familiarity and age ($p = .04$) and a significant three-way interaction between familiarity, age, and animacy ($p = .03$).

The significant two-way interaction between familiarity and age is illustrated in Figure 5. Children at around age three to five perform poorly, regardless of whether the items are nonce or familiar. From around age six to ten, they begin to produce more target classifiers for familiar than nonce items. This gap closes around 11 years old, where they perform near-ceiling on both familiar and nonce items. In sum, familiarity does not seem to play a role for younger and older children who perform at either floor or ceiling; but for those in middle childhood, it is more difficult to extend the classifier meaning to novice items in production than producing classifier-noun pairings that are available in their input.

This two-way interaction between age and familiarity is further modulated by animacy. As illustrated in Figure 6, the developmental differences between familiar and nonce items seem to be motivated from the animate items. When the items are animate, the developmental trajectories of familiar and nonce items differ (with familiar items developing faster) but when the items are inanimate, there are smaller differences in the rate of development between familiar and nonce items.

3.1.3 Conditional inference tree

The conditional inference tree, shown in Figure 7, corroborates the results of the previous mixed effects model, showing that age is a significant variable that predicts the production of target classifiers. The tree depicted in Figure 7 consists of a series of binary splits that divide the data into different subsets based on the predictor variables: age and familiarity (familiar and nonce). Each split represents a decision point where the tree branches into different paths. The tree selects the most relevant variables to create the splits based on their importance in predicting the outcome variable. The boxes at the bottom of Figure 7 represent the accuracy of classifier production from 0% to 100%. The first split occurs at age six, suggesting that age that is deemed most important for classifier production is six years old – children below this threshold have quite low accuracy while children above this threshold perform significantly better with more than 40% production accuracy (even for nonce items). Familiarity also plays a role in predicting their classifier production performance, especially for children who are under seven (see the bubble with
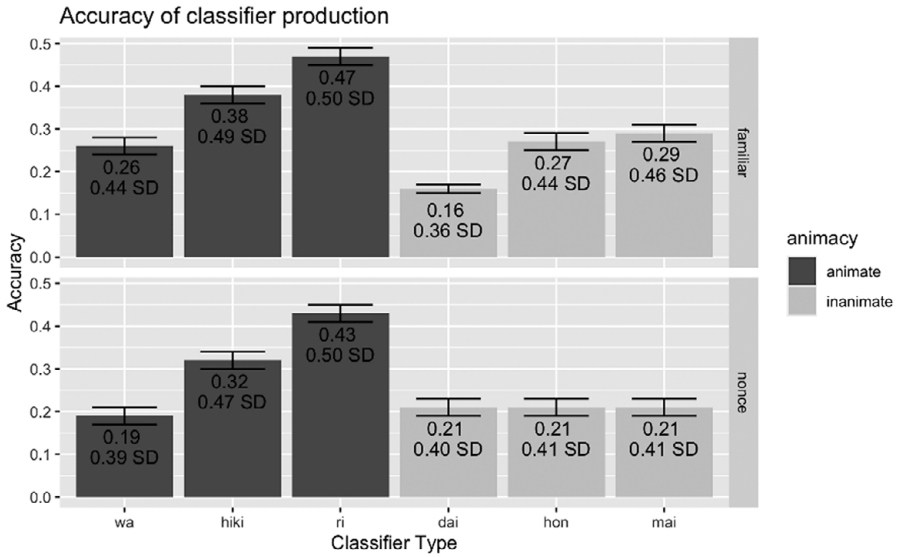
Figure 4. Accuracy of the classifier types split by animacy and familiarity for production. Error bars indicate standard error.

Table 3. The output of the generalized linear mixed effects model for production accuracy

Model: glmer(accuracy_production ~ familiarity * age *animacy+ (familiarity| participant) + (1| item))			
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.00	0.00 − 0.00	**<0.001**
familiarity [nonce]	2.94	0.51 − 17.01	0.229
age	4.27	3.02 − 6.05	**<0.001**
animacy [inanimate]	0.16	0.05 − 0.53	**0.003**
familiarity [nonce] * age	0.80	0.64 − 1.00	**0.047**
familiarity [nonce] * animacy [inanimate]	0.34	0.07 − 1.76	0.199
age * animacy [inanimate]	1.01	0.88 − 1.17	0.888
(familiarity [nonce] * age) * animacy [inanimate]	1.24	1.02 − 1.50	**0.031**
Observations		7878	
Marginal R^2 / Conditional R^2		0.404 / 0.845	

number 14 in Figure 7), with familiar items eliciting higher accuracy than nonce items. At seven years and above, children perform over 80% on both familiar and nonce items.

3.1.4 Qualitative analysis

Recall that the non-target responses (NTR) that the participants produced were coded within three types: (a) NC = no classifier; (b) GC = general classifier; (c) WC = wrong

Interaction between age and familiarity
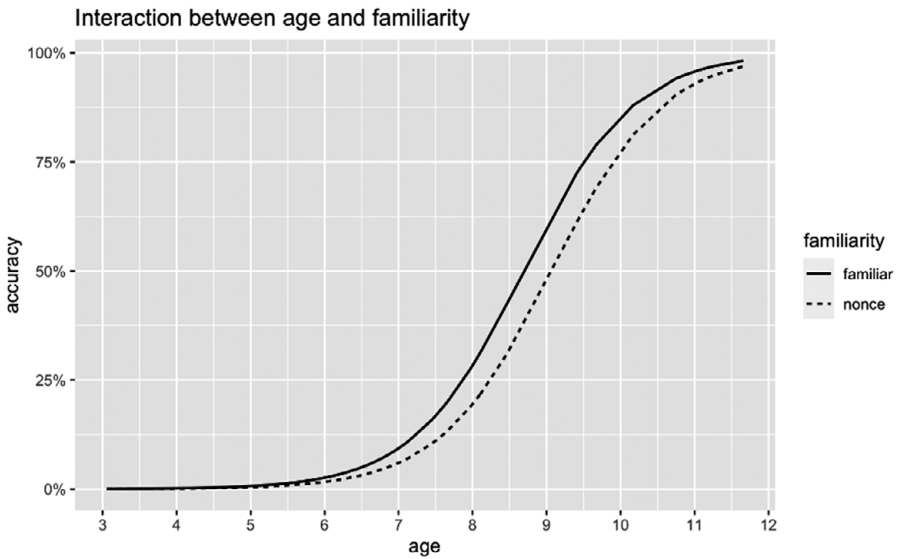
Figure 5. The plot of two-way interaction effects between age and familiarity.

Interaction between age, familiarity, and animacy
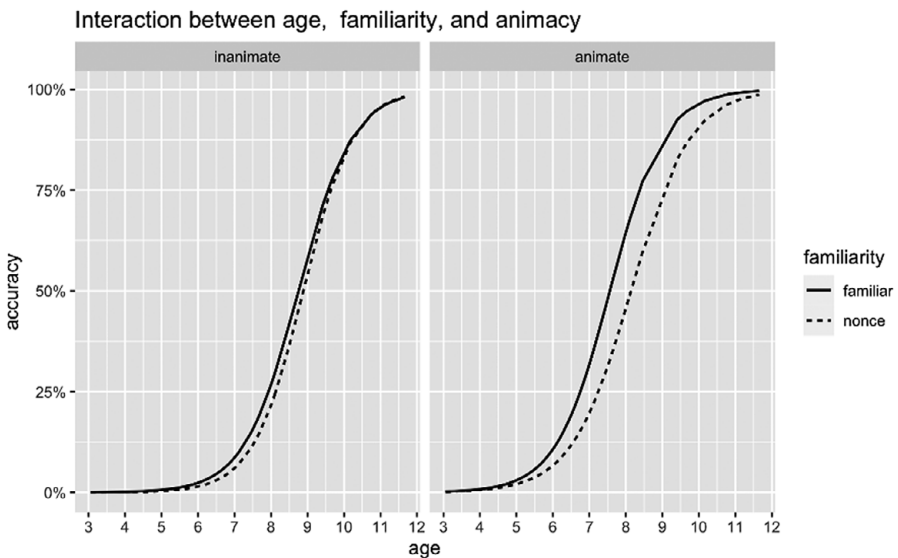
Figure 6. The plot of three-way interaction effects between age, familiarity, and animacy.

classifier. The descriptive statistics of the number of observations per type as a function of familiarity is provided in Table 4. NC was the most observed type of NTR (Sum = 2841, M = 12.25, SD = 6.49, 51% of all NTRs), followed by WC (Sum = 1391, M = 7.36, SD = 4.87, 25% of all NTRs) and GC (Sum = 1348, M = 9.30, SD = 5.63, 24% of all NTRs). Moreover, nonce items (Sum = 2858, M = 9.89, SD = 6.24) elicited slightly more NTR than familiar items (Sum = 2722, M = 9.83, SD = 6.05). Post-hoc comparisons using emmeans package
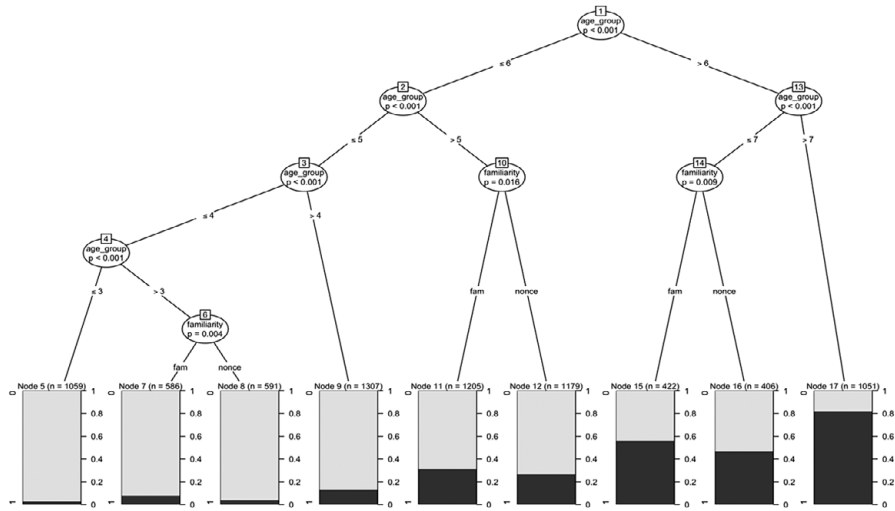
Figure 7. Conditional Inference Tree for production accuracy.

Table 4. Descriptive statistics of non-target response type for familiar and nonce items

Familiarity	Type	Number of obs	Percentage
familiar	GC	686	25.2%
familiar	NC	1376	50.5%
familiar	WC	660	24.2%
nonce	GC	662	23.1%
nonce	NC	1465	51.2%
nonce	WC	731	25.5%

with Tukey correction on the main effect of Type showed that occurrence of NC is greater than WC ($p = .009$) and GC ($p = .004$) and occurrence of GC was greater than WC ($p < .001$). The results of the poisson generalized linear mixed effects model are presented in Table 5. The only significant interaction was between familiarity and type ($E = 1.16$, $p = .03$). Post-hoc comparisons showed that the difference in the number of NTR between familiarity and nonce items varied across types, specifically between GC and NC ($E = .14$, $Z = 2.00$ $p = .04$). That is, there were more NTR for nonce items than familiar items for NC (i.e., no production of classifiers) when compared to GC (i.e., replacement of general classifiers).

As for the NC type, all classifiers elicited no classifiers/bare numerals to a similar extent (*wa* = 445; *hiki* = 459; *mai* = 474; *ri* = 475; *hon* = 480; *dai* =493). In terms of GC type, *-dai* elicited the most occurrences of general classifiers (n = 393) followed by *-hon* (n = 389) and *-mai* (n = 352). This is not surprising as the general classifiers (*-ko* and *-tsu*) can only

Table 5. The output of the poisson generalized linear mixed effects model for number of observations per non-target response (NTR) type

glmer(n ~ type*familiarity*age_z+ (type|participant), family=poisson)			
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	7.00	$5.66 - 8.66$	**<0.001**
type [NC]	0.85	$0.61 - 1.18$	0.332
type [WC]	0.78	$0.61 - 1.00$	0.050
familiarity [nonce]	0.93	$0.83 - 1.04$	0.185
age z	0.97	$0.80 - 1.18$	0.777
type [NC] * familiarity [nonce]	1.16	$1.01 - 1.34$	**0.037**
type [WC] * familiarity [nonce]	1.08	$0.93 - 1.27$	0.317
type [NC] * age z	0.76	$0.54 - 1.07$	0.117
type [WC] * age z	0.85	$0.67 - 1.08$	0.195
familiarity [nonce] * age z	0.97	$0.85 - 1.11$	0.661
(type [NC] * familiarity [nonce]) * age z	1.05	$0.89 - 1.24$	0.556
(type [WC] * familiarity [nonce]) * age z	1.08	$0.89 - 1.31$	0.430
Marginal R^2 / Conditional R^2	0.057 / 0.771		

Table 6. The ten most common non-target responses that were categorized as using the wrong classifier (WC)

Child's Response	Target Classifier	Number of obs
hiki	wa	214
too	hiki	151
too	wa	118
ri	hiki	76
hiki	ri	67
ri	wa	56
too	dai	39
too	mai	38
too	ri	35
hiki	mai	31

be used with inanimate objects. However, there were still some occurrences of the use of general classifiers for animate classifiers: *-ri* (n = 68), *-hiki* (n = 68), and *-wa* (n = 78). With respect to WC type, Table 6 illustrates the ten most common non-target responses that were categorized as using the wrong classifier. Since *-too* (classifier for large animals) was used in the example video which explained the task to the children, many of the responses

categorized as WC involved the use of *-too* classifier for objects that take on animate classifiers (*-ri, -hiki, -wa*) but also for inanimate classifiers (*-dai, -mai*).

3.2. Comprehension task

3.2.1 Descriptive results

The accuracy and reaction time of the comprehension of classifier types split by animacy and familiarity are presented in Figure 8. Interestingly, in contrast to the production results,
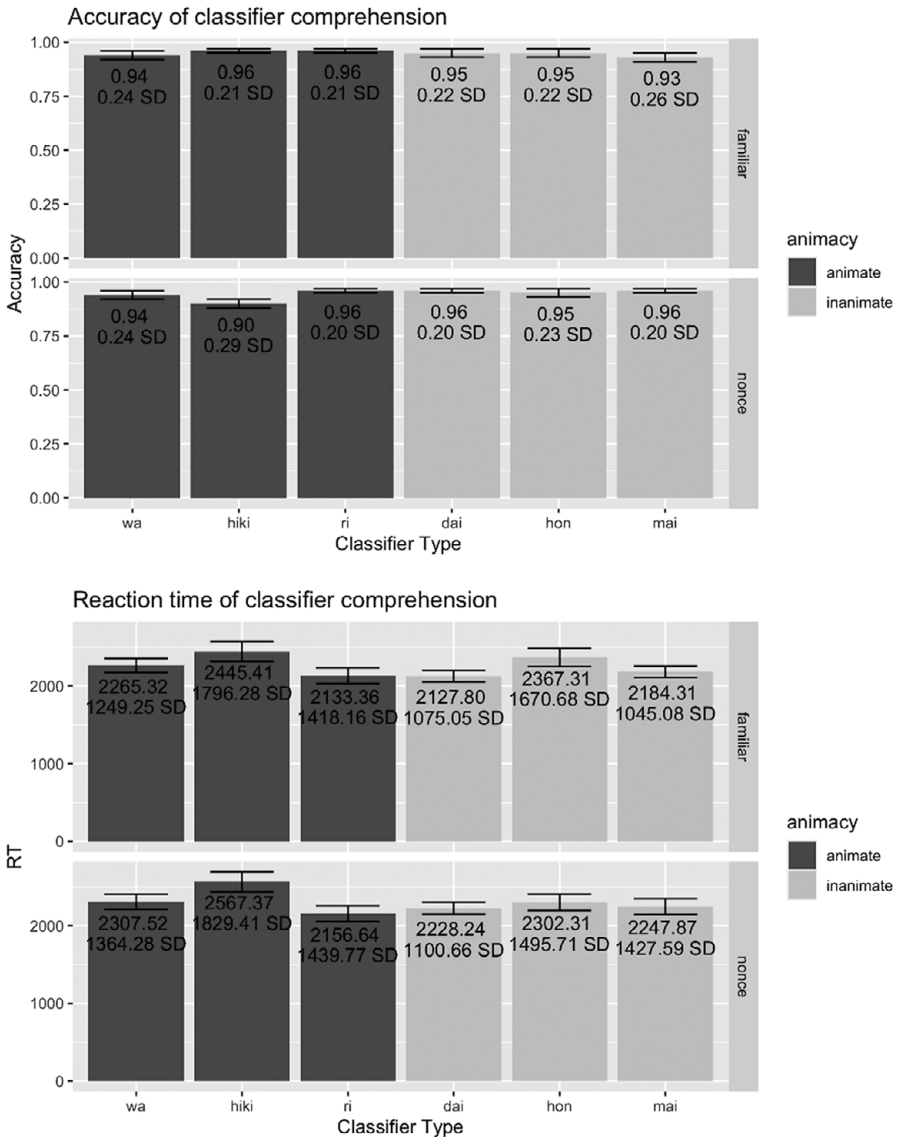

Figure 8. Accuracy and reaction time of the classifier types split by animacy and familiarity for comprehension. Error bars indicate standard error.

children performed at ceiling (more than 90%) on all classifier types, regardless of whether the item was animate (Accuracy: $M = .94$ $SD = .23$, RT: $M = 2312.58$, $SD = 1536.51$) or inanimate (Accuracy: $M = .95$, $SD = .22$, RT: $M = 2242.96$ $SD = 1323.77$), or familiar (Accuracy: $M = .95$, $SD = .22$, RT: $M = 2254.02$ $SD = 1406.84$) or nonce (Accuracy: $M = .95$, $SD = .23$, RT: $M = 2301.48$, $SD = 1461.17$). The accuracy and reaction time of classifier comprehension split by each age group, animacy, and familiarity can be found in the Supplementary Materials.

3.2.2 Generalized and linear mixed effects model

The summary outputs of the generalized linear mixed effects model (Accuracy) and linear mixed effects model (log RT) are presented in Table 7. The only significant predictor in both models was age (Accuracy: $p < .001$, logRT = $p < .001$), and there was no significant interaction between age and familiarity (p's > .53) or a three-way interaction between age, animacy, and familiarity (p's > .06). These results demonstrate that children become more accurate and faster in comprehending classifiers as they grow older, but this developmental trajectory does not differ between familiar/nonce or inanimate/animate items.

3.2.3 Conditional inference tree

The conditional inference tree of the comprehension accuracy is presented in Figure 9. Here, we can see that the first split occurs at four years old, suggesting that this is a crucial age in which development of comprehension of classifiers takes place. Unlike in the production task in which familiarity influenced the children's performance until age seven, this factor only played a role in comprehension performance among children who are three years old (see bubble marked with number 3), with familiar items eliciting higher accuracy than nonce items. These findings suggest that comprehension of common classifiers, including its semantic information, is acquired by the age of four in monolingual Japanese children.

3.3. Mouse-tracking results

Mean mouse trajectories are visualized in Figure 10 as a function of familiarity and animacy. Visual inspection and descriptive results indicate that mouse movements do not diverge between familiar (MAD = .41, Entropy = .11) and nonce items (MAD = .40, Entropy = .11) or between animate (MAD = .39, Entropy = .11) and inanimate items (MAD = .41, Entropy = .11). This is also supported by the output of the linear mixed effects model that we ran separately for MAD and Entropy. In terms of the MAD model, the only significant predictor was age ($E = -.05$, $t = 4.52$, $p < .001$), and no other main effects or interactions were significant (p's > .13). For the sample entropy model, no significant main effects or interactions were found (p's > .47). In sum, the mouse-tracking results show that children perform similarly regardless of the familiarity or the animacy of the item, suggesting that they are able to process the semantic properties of classifiers in real-time and generalize them to novice contexts.

Table 7. The output of the generalized linear mixed effects model (Accuracy) and linear mixed effects model (RTs) for comprehension

Model: glmer/lmer(accuracy/log(RT) ~ familiarity * age *animacy+ (familiarity| participant) + (1| item))						
	Accuracy			Log RT		
Predictors	Odds Ratios	CI	p	Estimates	CI	p
(Intercept)	0.05	0.01 – 0.18	**<0.001**	8.82	8.64 – 9.01	**<0.001**
age	2.90	2.29 – 3.68	**<0.001**	-0.13	-0.15 – -0.10	**<0.001**
familiarity [nonce]	0.69	0.25 – 1.89	0.468	0.08	-0.02 – 0.19	0.115
animacy [inanimate]	0.81	0.28 – 2.39	0.706	0.04	-0.07 – 0.15	0.453
age * familiarity [nonce]	0.95	0.77 – 1.18	0.667	-0.01	-0.02 – 0.01	0.532
age * animacy [inanimate]	0.98	0.79 – 1.22	0.865	-0.01	-0.02 – 0.01	0.437
familiarity [nonce] * animacy [inanimate]	0.57	0.13 – 2.49	0.458	-0.04	-0.17 – 0.09	0.554
(age * familiarity [nonce]) * animacy [inanimate]	1.34	0.98 – 1.83	0.066	-0.00	-0.02 – 0.02	0.927
Observations	5789			5789		
Marginal R^2 / Conditional R^2	0.538 / 0.675			0.221 / 0.496		
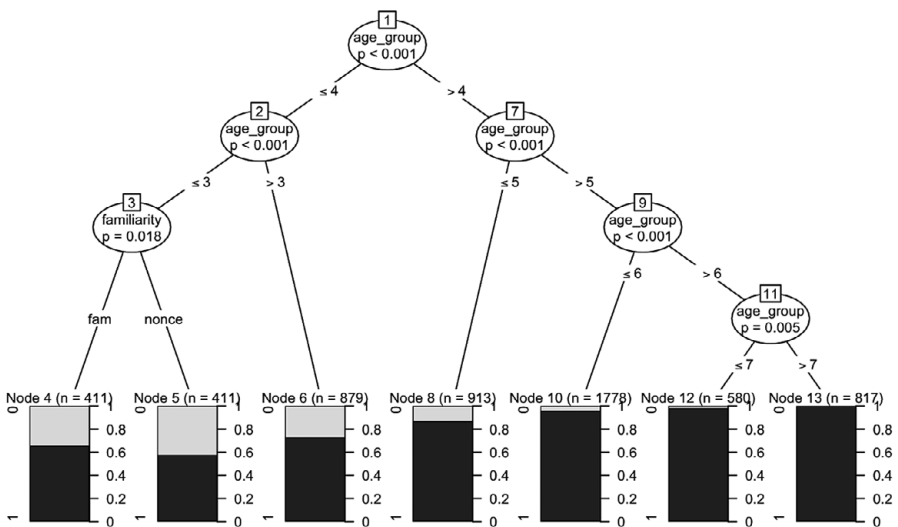
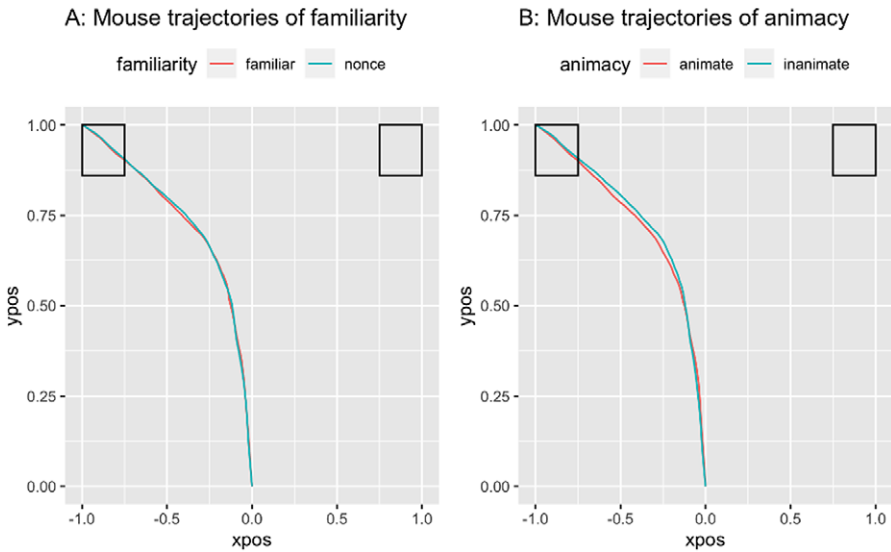**Figure 9.** Conditional Inference Tree for comprehension accuracy.

Figure 10. Mean mouse trajectories split by familiarity (Panel A) and animacy (Panel B).

4. Discussion

The current study investigated the production, comprehension, and processing of classifiers using familiar and nonce items in 3 to 12-year-old Japanese monolingual children. Not least because existing evidence is mixed and highly skewed towards production, the primary goal of combining modalities was to examine THE TIMING of the acquisition of the semantic system of basic numeral classifiers in childhood from converging evidence bases. The present study is set up to also address WHETHER, and if so WHY, evidence for acquisition depends on the modality of testing. In order to examine children's ability to extract semantic information of classifiers and extend it to novel situations, our experiments combine familiar and nonce stimuli to ensure that the participants have never encountered such classifier-noun pairings in the input.

To sum up the PRODUCTION data, results show that children perform better on familiar than nonce items, but this pattern was significantly modulated by age. That is, for younger or older children who performed at floor or at ceiling respectively, their performance did not differ, regardless of whether the items were familiar or nonce. However, for those in middle childhood who are in the process of developing their classifier knowledge, having a form-class cue in the input correlates to increasing accuracy in the production of expected classifiers. Not surprisingly given its prominence in the Japanese system, this pattern was further modulated by animacy – children as young as five showed increased production accuracy for familiar over nonce items when these items were animate – providing independent evidence that semantic features can drive production choices, even before ceiling levels are reached. However, this pattern is asymmetric: a difference between inanimate familiar and nonce items is never attested at any age, irrespective of overall accuracy increasing as a function of age. Given that inanimate nouns potentially have less cognitive and/or pragmatic salience, at the youngest of ages our child participants

(perhaps Japanese children in general) may not have had enough experience (or cognitive resources) to bootstrap the production of classifiers with inanimate items.

The output of the Conditional Inference Tree indicates that six years of age is a critical point in which development of classifiers takes place for PRODUCTION, which most likely coincides with introduction to formal schooling and enhancement in literacy skills, as they receive more quantitatively and qualitatively rich input from the environment. Most importantly, children do not perform equally well on familiar and nonce items until they are seven years old. Our results are in similar vein to what has been suggested previously in the literature (Uchida & Imai, 1996, 1999). If the present study only had production data, we might be inclined to argue in favor of significantly protracted development of the semantics of numeral classifiers. Although possible, we argue that converging evidence beyond production in isolation is needed to make such a conclusion. While we do acknowledge the importance of production data in acquisition studies, production alone cannot be conflated with, as if being equivalent to, acquisition proper. As the results above indicate (which we will proceed to unpack below), both the comprehension and processing parts of the study rather suggest that the production evidence underdetermines the system that younger Japanese children have. In other words, while we replicate the patterns that have previously been shown in production studies, all that can be definitively concluded from the present production data is that Japanese children have protracted PRODUCTIVE development of classifiers in relation to their semantics, not the representation of the basic semantic system of the classifiers *per se*.

Before delving deeper into the complementary modalities of testing, it is worth commenting on a particular aspect of the qualitative analysis for production, as the pattern departs from previous studies. In summary, the most common non-target responses made by children are omissions of classifiers (NC; counting items with bare numerals), followed by substitutions with general classifiers (GC) and, then, use of incorrect classifiers (WC). This finding is in contrast with other studies finding omissions of classifiers to be rare even among younger Japanese and Chinese children (Hao et al., 2021; Uchida & Imai, 1999). Such discrepancies may be affected by the inclusion of nonce items in this study. Encountering nonce items may have influenced the children to opt for a strategy of completely omitting classifiers when they could not extend the meaning of classifiers to novice contexts. This is indexed by the fact that children omitted classifiers more frequently when counting nonce items than familiar ones. Out of 120 participants, 20 children did not produce any classifiers (16% of the population, average age = 4.80, SD = 1.15) and 35% of children's total production (both correct and incorrect) consisted of classifier omission responses. While our study shows a comparatively higher rate of omission, it does not deviate too greatly from what Uchida and Imai (1999) found in their study using only familiar items, in which they found that Japanese monolingual four-year-olds make classifier omission errors around 20% of the time.

The findings of the COMPREHENSION task were in stark contrast to the production results – children's accuracy was above 90% on all classifier types and there was no difference in accuracy between familiar vs. nonce or animate vs. inanimate items. Moreover, there was no interaction between age, familiarity, or animacy for both reaction time and accuracy. This suggests that the acquisition of the underlying semantics of common classifiers is robustly represented from the earliest of ages tested, such that in comprehension it can be deployed equally for familiar and nonce items, regardless of animacy (or the target-competitor pairings as indicated in the Supplementary Materials). The only significant predictor for comprehension was age. That is, children became more accurate and faster in comprehending classifiers as they grew older. The significant role of

age is also found in processing via the mouse-tracking results, with no interactions between familiarity or animacy for measurements of movement trajectory (MAD) and complexity (sample entropy). This demonstrates that children become better at processing classifiers in real-time as they get older, and they are also able to process the semantic system of classifiers at a relatively young age.

The output of the Conditional Inference Tree corroborates the behavioral and mouse-tracking results, indicating that significant improvement in COMPREHENSION of classifiers takes place around the age of four. This is in line with what Uchida and Imai (1999) and Yamamoto and Keil (2000) found in their study of Japanese monolingual children using comprehension tasks. Acquisition of the common classifiers, therefore, seems to take place at least two years prior to production (which occurs around the age of six). In addition, while familiar items had higher accuracy than nonce items until age seven for production, only three-year olds displayed such differences for comprehension. This suggests that at least by the age of three to four, children have acquired the semantic properties of the most common classifiers for comprehension and are able to extend it to novel items with high accuracy.

Given that production appears to lag behind comprehension and processing for several years, a natural question emerges: Why is there such a gap? Yamamoto and Keil (2000) speculate that this is due to several converging factors such as: (a) the form of classifiers changing depending on the preceding numerals (e.g., human classifiers: hito-*ri*, futa-*ri*, sann-*nin*); (b) alternation of the phonological shape depending on the preceding numeral (e.g., long, thin classifiers: ip-*pon*, ni-*hon*, san-*bon*); (c) the use of incorrect classifiers not hindering communication (if presented with a noun). However, such factors do not apply only to production – indeed, successful comprehension of classifiers also requires children to understand that the classifier form and phonology changes depending on the preceding numeral. Additionally, contrary to what Yamamoto and Keil (2000) suggest, comprehension of the grammatical or semantic function of classifiers is not necessarily needed for production than it would be relevant for comprehension, as long as children can understand the properties of numerals and nouns (e.g., when a child hears "inu ga ip-piki (dog-particle-one-CL)", they do not necessarily need to understand the function of the classifier to interpret that the speaker is talking about one dog). Rather, we stipulate that the discrepancy stems from processing issues related to activation of grammatical and semantic information of classifiers. After all, there is more involved in production (an active process) than comprehension and processing (a comparatively more passive one) such that one can demonstrate empirically "knowing more than they say" (González Alonso & Puig-Mayenco, 2021; Hendriks & Koster, 2010).

Let us consider a few aspects implicit to differences in production and comprehension/processing methods that might further illuminate the present asymmetry. In comprehension experiments such as the ones we used, the child is provided with a specific classifier (e.g., hon) and they have to first decode its semantic function by extracting its specific features (e.g., long, thin, 3D). The next step involves matching the extracted semantics to the inherent features of the target item – or in the current experimental paradigm, choosing a picture out of two options that best matches the classifier features to those of specific items. In other words, they primarily have to decode (and match) something that is given to them.

In production, a similar process (this time decoding from a picture), albeit in reverse, applies. However, in doing so the ultimate task of producing language is not accomplished: encoding must also take place precisely because the child is not afforded the specific classifier in the experimentation. Alternatively, on the basis of a picture, the child

first needs to access the lexical entry for the target item noun, inclusive of its relevant semantic features, to then also activate and access the associated classifier (e.g., shape, color, animacy, texture, function etc.). In doing so, they have to narrow down the features to the ones that best represent the item (e.g., long, thin, brown, green, inanimate. 3D), while activating the features of potential classifiers (e.g., mai = thin, flat, 2D; hon = thin, long, 3D) and, only then, selecting/producing a classifier in which the features match. Given this, there is a potential for more optionality implicit to the production of classifiers, a context which should then impart greater processing demands.

From what we have seen in the sum of the present data across methods, the asymmetry suggests at least a few things. Firstly, specific classifiers are indeed not stored as part of the lexical inventory of particular nouns. Rather Japanese classifier selection embodies the underlying matching of semantic features across related, but, crucially, distinct lexical items. Secondly, the semantic system is not problematic for acquisition *per se*, but rather the matching of features between the two lexical items required by the grammar is the issue. There is an accessing cost that is not fully overcome until relatively late in acquisition/cognitive developmental terms, lingering in later childhood when the activity is more demanding (i.e., surfacing in production). In other words, while the development of productive ability of classifiers is more protracted than comprehension, comprehension data show that children are able to extend semantic properties of classifiers to novice contexts as young as age three.

5. Conclusion

In sum, our findings show that, distinct from what has been argued/assumed in previous literature, children as young as three acquire the semantic system of basic Japanese classifiers. We argue this on the basis of convergent data from two types of comprehension, offline behavior and online processing. Crucially, not only do children perform well in comprehension of Japanese classifiers overall, a pattern previously attested (Sumiya & Colunga, 2006; Uchida & Imai, 1996, 1999; Yamamoto & Keil, 2000), by juxtaposing comprehension and production with the same participants and by incorporating nonce words in our experiments, we can offer firmer conclusions. While previous work (Uchida & Imai, 1999) also found a similar asymmetry, because they only used familiar words, it is not immediately clear from their data that the semantic system has been acquired. Alternatively, it could be the case that their children have associated specific classifiers to particular known nouns (lexically), the accessing of which is somehow less problematic/costly in comprehension over production. Given the fact that the present comprehension data show no asymmetry between familiar and nonce items (or any influences from target-competitor pairings), it must be the case that young Japanese children have acquired the underlying semantics of the classifier system, at least in terms of the most common classifiers. Absent of this, one leaves unexplained how they were able to extend such knowledge for items encountered for the first time in the experiments (and crucially, in line with the norming study we ran with adult Japanese speakers).

Replicating results from previous production-based methods, the present data also evidence a delay in achieving adult-like mastery of classifier selection in production. We speculated that the developmental asymmetry between comprehension and production is due to greater task complexity/processing demands involved in production. Because children have limited cognitive resources to handle the greater processing demands, having both semantic information and form-class cue (instead of semantic information

alone) mitigates the processing load, contributing to higher accuracy of classifiers for familiar over nonce items – a finding that could only be revealed and whose significance understood in a method such as the one adopted here. However, we would like to highlight that the focus on the present case does not imply a universally superior role for comprehension over production. Instead, it demonstrates that, in this particular instance, it is justifiable to propose that comprehension data from young children indicate a level of knowledge that has not been extensively examined in prior related research, which has primarily focused on production. Various factors may contribute to the asymmetries in performance between comprehension and production, and the answer seems to lie in aspects of production that fall beyond the scope of what comprehension can assess – specifically, outside the realm of underlying grammatical representation itself. If we are on the right track, future studies should further examine how domain-general cognitive ability interacts with the developmental trajectories of Japanese classifiers, at the aggregated and individual levels, specifically examining the question of whether cognitive ability is an important predictor for classifier production.

References

Adams, K. L., & **Faires Conklin, N.** (1973). Toward a theory of natural classification. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. **9**, No. 1, pp. 1–10). Chicago Linguistic Society.

Aikhenvald, A. Y. (2000). *Classifiers: A typology of noun categorization devices*. OUP Oxford.

Bates, D., **Kliegl, R.**, **Vasishth, S.**, & **Baayen, H.** (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Chien, Y. C., **Lust, B.**, & **Chiang, C. P.** (2003). Chinese children's comprehension of count-classifiers and mass-classifiers. *Journal of East Asian Linguistics*, **12**(2), 91–120.

Downing, P. (1996). *Numeral Classifier System: The Case of Japanese*. John Benjamins: Amsterdam.

González Alonso, J., & **Puig-Mayenco, E.** (2021). You know more than you say: Methodological choices in L3 transfer research. *Linguistic Approaches to Bilingualism*, **11**(1), 54–59.

Hao, Y., **Bedore, L.**, **Sheng, L.**, **Zhou, P.**, & **Zheng, L.** (2021). Exploring influential factors of shape classifier comprehension and production in Mandarin-speaking children. *First Language*, **41**(5), 573–604.

Hehman, E., **Stolier, R. M.**, & **Freeman, J. B.** (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, **18**(3), 384–401.

Hendriks, P., & **Koster, C.** (2010). Production/comprehension asymmetries in language acquisition. *Lingua*, **120**(8), 1887–1897.

Hothorn, T., **Hornik, K.**, & **Zeileis, A.** (2015). ctree: Conditional inference trees. *The comprehensive R archive network*, **8**.

Jarkey, N., & **Komatsu, H.** (2019). Numeral classifiers in Japanese. In A. Y. Aikhenvald & E. Mihas (Eds.), *Genders and Classifiers: A Cross-Linguistic Typology, Explorations in Linguistic Typology* (pp. 344–395). Oxford, UK: Oxford University Press.

Kieslich, P. J., **Henninger, F.**, **Wulff, D. U.**, **Haslbeck, J. M.**, & **Schulte-Mecklenbeck, M.** (2019). Mouse-tracking: A practical guide to implementation and analysis 1. In *A handbook of process tracing methods* (pp. 111–130). Routledge.

Levshina, N. (2015). *How to do linguistics with R. Data Exploration and Statistical Analysis*. John Benjamins.

Li, P., **Barner, D.**, & **Huang, B. H.** (2008). Classifiers as count syntax: Individuation and measurement in the acquisition of Mandarin Chinese. *Language Learning and Development*, **4**(4), 249–290.

Li, P., Huang, B., & Hsiao, Y. (2010). Learning that classifiers count: Mandarin-speaking children's acquisition of sortal and mensural classifiers. *Journal of East Asian Linguistics*, **19**(3), 207–230.

Matsumoto, Y. (1987). Order of Acquisition in the Lexicon: Implication from Japanese Numeral Classifiers. In K. E. Nelson and A. Kleeck (eds.), *Children's Language*, 7, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 229–260.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Salehuddin, K., & Winskel, H. (2009). An investigation into Malay numeral classifier acquisition through an elicited production task. *First Language*, **29**(3), 289–311.

Sanches, M. (1977). Language acquisition and language change: Japanese numeral classifiers. In M. Snaches and B. Blount (eds.), *Sociocultural dimensions of language change.* Academic Press, NY, pp.51–62.

Sera, M. D., Johnson, K. R., & Kuo, J. Y. (2013). Classifiers augment and maintain shape-based categorization in Mandarin speakers. *Language and Cognition*, **5**(1), 1–23.

Sumiya, H., & Colunga, E. (2006). The effect of familiarity and semantics on early acquisition of Japanese numerical classifiers. *Proceedings of the 30th Annual Boston University Conference on Language Development*, 607–618.

Uchida, N., & Imai, M. (1996). A study on the acquisition of numerical classifiers among young children: The development of human/animal categories and generation of the rule of classifiers applying. *Japanese Journal of Educational Psychology*, **44**(2), 126–135.

Uchida, N., & Imai, M. (1999). Heuristics in learning classifiers: The acquisition of the classifier system and its implications for the nature of lexical acquisition. *Japanese Psychological Research*, **41**(1), 50–69.

Yamamoto, K., & Keil, F. (2000). The acquisition of Japanese numeral classifiers: Linkage between grammatical forms and conceptual categories. *Journal of East Asian Linguistics*, **9**(4), 379–409.