

Abstract

The current study examined the comprehension and production of classifiers, case-marking, and morphological passive structures among 414 child Japanese heritage speakers (Mean age = 10.01; Range = 4.02 – 18.18). Focusing on individual differences, we extracted latent experiential factors via the Q-bex questionnaire (DeCat. et al., 2022) which were then used to predict knowledge and use of these grammatical structures. The findings reveal that: (i) experiential factors such as heritage language (HL) engagement at home and within community modulate grammatical performance differentially from childhood through adolescence and (ii) HL proficiency, immersion experiences, and literacy systematically predict HL grammatical outcomes. These results indicate that particular language background factors hold differential significance at distinct developmental stages and that higher proficiency, richer immersion experiences, and literacy engagement in the HL are crucial for the development of core grammatical structures.

1. Introduction

Bilingual children exhibit greater diversity in their language learning environments and individual language outcomes compared to monolingual children. This variation highlights the significance of adopting an individual differences approach in studying bilingual development; examining the sources of variability on linguistic and psycholinguistic measures rather than examining group comparisons alone (Paradis, 2023; Rothman et al., 2023). In recent years, such an approach has increasingly been adopted in heritage language bilingualism (HLB), a situation where individuals grow up in a community or family environment where a language other than the majority or dominant language of the larger society is spoken (Montrul, 2008; Rothman, 2009). Unpacking the relative contribution of factors that give rise to the significant variability observed in HL development and outcomes—across groups and within the individuals comprising them—is important beyond its manifold theoretical relevance. Doing so is crucial for parents, clinicians, and educators to support children’s HL development via family practices, educational programs, and interventions.

Morphosyntactic studies in child HLB have examined a wide range of grammatical structures ranging from word order (Hao & Chondrogianni, 2021; van Osch et al., 2019), case-marking (Chondrogianni & Schwartz, 2020; Meir & Janssen, 2021), passives (Bayram et al., 2019), gender (Mitrofanova et al., 2022; Rodina et al., 2020), and classifiers (Kan, 2019) and more. This work underscores that various factors modulate HL grammars such as parental input quality (Daskalaki et al., 2020), age of onset (Jia & Paradis, 2015; Soto-Corminas et al., 2022), home language use (Flores et al., 2017; Kan, 2019), cumulative length of exposure (Mitrofanova et al., 2018), literacy/formal education (Bayram et al., 2019; Torregrossa et al., 2023) and immersion experiences (Chondrogianni & Daskalaki, 2023) among others. Moreover, these studies demonstrate that language domains can be

differentially affected by child-external factors, especially quality and quantity of input; whereby more input seems to be necessary for vocabulary than grammatical structures (Chondrogianni, 2023). It can also be modulated by the internal structure of the language—for example, syntax-discourse interface structures may be more difficult to acquire/process and thus require more input than (narrow) syntactic ones (Sorace, 2011).

Although significant advances have been made in terms of uncovering the sources of individual variability in heritage speakers (HSs), the vast majority of previous studies are small to mid-scaled (i.e., participant numbers under 100, especially those investigating beyond vocabulary) and examine specific age ranges by regressing single questionnaire responses (or taking the average of multiple responses) to HSs' linguistic competence/performance. However, it is highly probable that various factors explain HL grammars differentially from childhood through adolescence and/or simply as a function of specific community/individual context. For instance, HL exposure and use at home may be a crucial variable for HL grammar until HSs enter formal education, but after this transition other factors such as HL engagement in the society and school, or literacy may better explain the developmental trajectory and possibly their eventual outcomes of specific domains. Nevertheless, previous studies (perhaps because of their limited age range and participant numbers) have primarily investigated whether factor X predicts grammar Y, rather than accounting for the interactions between factor X and chronological age on Y (if it is a cross-sectional design) or examining the effect of factor X on Y longitudinally from childhood through adolescence. Some exceptions come from recent studies by Torregrossa et al. (2023) that have examined the acquisition of several grammatical structures in European Portuguese as a HL among 180 HSs from age 8 to 16. They found that HS children's linguistic competence generally increased with age, and formal instruction in the HL correlated with higher accuracy. Another study by Daskalaki et al. (2022) tested subject placements and

subject/object forms in 61 Greek-English HS children (age 6.5 to 19) and their monolingual peers. They found that accuracy on some conditions (wide focus) increased with age while others (topic continuity) did not, and HL input and generation were consistent factors that predicted their grammatical competence. Moreover, in a one-year longitudinal study that tracked morphosyntactic development in Arabic-English HL children from age 6 to 13, Paradis et al. (2021) showed that cognitive skills predict morphosyntactic competence over time in both languages, while age of L2 onset and input factors differentially affected the HL and the majority language longitudinally.

Our large-scale study with over 400 Japanese HSs ranging from age four to eighteen allows us to zoom in on this issue by investigating comprehension and production of structures that have been shown to be particularly vulnerable in the HLB literature (classifiers and case-marking/passives). Additionally, we employ a newly developed questionnaire, Q-bex (De Cat et al., 2023), which is specifically designed to explore in detail the dual language experiences/engagement and traits of child bilingualism. We utilize Exploratory Factor Analysis (EFA) to extract latent factors that represent the underlying structure of the questionnaire responses and interact them with (chronological) age in our regression analyses to examine when and how different factors predict various HL grammars within the developmental trajectories of HSs. Moreover, we further examine whether this effect is modulated by semantic and (morpho)syntactic manipulations such as familiarity (nonce, familiar), canonicity (canonical SOV, non-canonical OSV), and voice (active, passive). To this end, in the following sections, we will briefly describe the properties of the target structures before stating our research questions.

1.1 Classifiers

Classifiers are special expressions accompanying numerals to categorize items (nouns) that they quantify and are common in East Asian languages like Korean, Mandarin,

and Japanese (e.g., Japanese: *san too no raion* ‘three CL-GEN lion’). They are used to specify the type of objects being counted and often reflect certain inherent characteristics of those objects. In Japanese, classifiers are an essential aspect of its grammatical structure and must be attached to a noun whenever the quantity is specified. While some universal principles, like animacy and shape, apply to various classifier languages, there are complexities and differences unique to each. In Japanese, classifiers are strictly divided into two main categories: animate and inanimate and are further organized based on semantic features like animal type, shape, and function as illustrated in Figure 1 (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. Moreover, 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 cars).

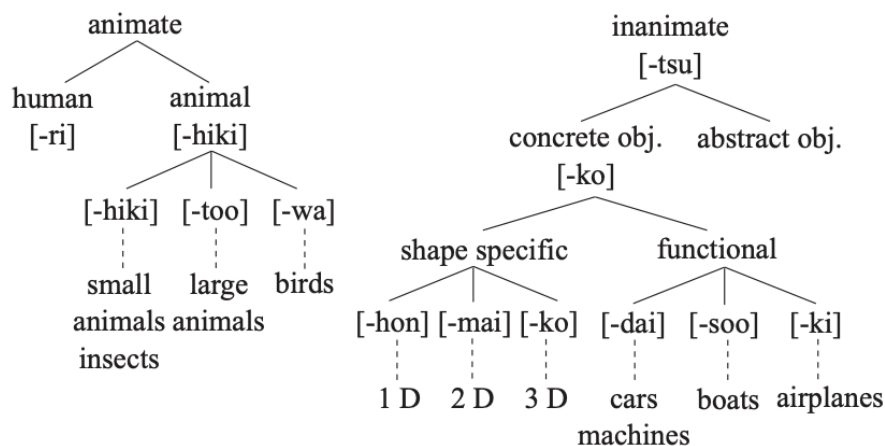


Figure 1. Japanese numeral classifier system (taken from Yamamoto & Keil, 2000, p.381)

A central question regarding the acquisition of classifiers is whether children acquire not only the syntactic function of the classifier (i.e., mapping classifier to the noun) but also the underlying semantic rules associated with specific classifiers (e.g., *-hon* is a classifier for

long thin objects). Recent studies using nonce and familiar items to test classifier knowledge among monolingual Japanese (Kubota et al., 2024) and Mandarin (Li et al., 2010) children suggest that children as young as three perform on par with familiar and nonce items, indicating they are able to extract semantic information and extend it to novel situations at a relatively young age. However, it is expected that even monolingual children require ample exposure to exemplars from each classifier category to learn to apply semantic rules to entities that share similar properties (Uchida & Imai, 1999). Thus, the acquisition of classifiers may be a domain of grammar affected more than others with respect to the quantity and quality of exposure. As such, it may display greater variability across HS individuals relative to their experience and engagement with Japanese.

Although limited, there is some evidence that shows vulnerability in the production of classifiers in HSs (Kan, 2019; Ruiting, 2016). Kan (2019) demonstrated that Cantonese HSs were less accurate in selecting the appropriate classifier than monolingual peers and showed a wide range of variability in their performance. Ruiting (2016) found that HSs of Mandarin performed differently from their monolingual counterparts in terms of their knowledge of classifiers, while they converged on tasks measuring their syntactic knowledge such as relative clauses and post-verbal clauses. Moreover, other studies have demonstrated a link between parental input and acquisition of classifiers (Atagi & Sandhofer, 2015; Naka, 1999). For instance, Atagi and Sandhofer (2015) reported that there was a high correlation between parents' and children's frequency of classifier use. In addition, the number of specific types of classifiers used by parents was also highly associated with the number of specific types of classifiers used by children. Finally, Hao et al., (2024) investigated the processing of classifiers in Mandarin Chinese heritage speakers via eye-tracking and found that home language use and formal schooling in the HL modulated their predictive processing of classifiers.

1.2 Case-marking and passive structures

Noun phrases are typically case marked in Japanese by a post-positional case marker, although it can be dropped in some cases. As illustrated in the SOV ordered sentence in (1), the case marker *-ga* indicates the nominative ‘chicken’, while the case marker *-o* shows the accusative ‘dog’. As can be seen in sentence (2), Japanese allows constituents to occur in a variety of orders (OSV) with the same meaning as in (1) whereby syntactic function is clearly reflected by case morphology.

(1) niwatori-ga inu-o osu
 chicken-NOM dog-ACC push
 ‘A chicken pushes a dog’

(2) inu-o niwatori-ga osu
 dog-ACC chicken-NOM push
 ‘A chicken pushes a dog’

As shown in (3) below, Japanese direct passives allow logical objects to appear as the grammatical subject and the logical subject as the “by-phrase” which is marked with a particle *-ni*. Moreover, it encodes the passive voice with the morpheme—*(ra)re*—on the verb. Due to the fact that Japanese is a head-final language, the verb—which allows for disambiguating information syntactically—does not appear until the end of the sentence. Furthermore, the case-marking morphemes *-ga* (nominative) and *-ni* (dative, locative/instrumental, or by-agent) can appear in non-canonical sentences such as example (4) but employs the same meaning as the canonical structure in (3).

(3) niwatori-ga inu-ni o-sareru
 chicken-NOM dog-DAT push-PASS
 ‘A chicken is pushed by a dog’

(4) inu-ni niwatori-ga o-sareru
 dog-DAT chicken-NOM push-PASS
 ‘A chicken is pushed by a dog’

In terms of case-marking, some studies have found that Japanese monolingual children do not produce the subject marking *-ga* like adults until at least the age of three (Noji, 1985), while others argue that Japanese children start understanding single-argument sentences with case markers *-ga* and *-o* around the age of four, but their performance does not reach an adult-like level until around the age of five to six (Suzuki, 2005). It has also been shown that young adult Japanese HSs exhibit high rates of *-o* omission and overuse of *-ga* (Laleko & Polinsky, 2013) and Korean HSs (Korean case-marking of SOV and OSV structures functions similarly to Japanese; Laleko & Polinsky, 2013) display poor comprehension of OSV/non-canonical word order suggesting that case-marking and particles can be vulnerable properties in HLB.

Passives are also a property that has been shown to be vulnerable in HL grammars, the degree of which correlates to experiential factors. For instance, Hao et al., (2021, 2023) examined Chinese passives (i.e., *bei*-constructions) in child HSs, showing they displayed less target-like performance compared to their monolingual peers on offline comprehension and production. Another study by Bayram et al., (2019) found that production of passives in Turkish HSs was modulated by the amount of literacy/formal training in the HL. Taken together, both case-marking and passive structures are good candidates to test in the context of our study, given their potential variability in competence and performance across different HS groups/individuals.

1.3 Research Questions

In light of the above, the research questions are formulated as follows:

- 1) What underlying experiential factors (extracted from the Q-bex questionnaire) predict the development of comprehension and production of classifiers, case-marking (in active and passive voice) and morphological passive structures in Japanese child HSs?

- 2) Do various experiential factors modulate grammatical performance differentially from childhood through adolescence? If so, are these effects further modulated by semantic and (morpho)syntactic manipulations such as familiarity (in case of classifiers), canonicity, and voice?

2. Methods

2.1 Participants

We initially collected data from 457 HS participants. However, we excluded those (a) not meeting the age criteria (see below) (b) indicating language impairment or developmental disorders (c) with missing questionnaire or linguistic data (d) reporting not having been exposed to Japanese before the age of three (e) not residing in a non-Japanese majority language (ML) context for at least two-thirds of their life. According to these criteria, 41 participants were excluded. Thus, the final number of participants in the study was 414 Japanese HSs (Mean age = 10.01; Range = 4.02 – 18.18; Female = 206; 304 from English-dominant environment, 110 from German-dominant environment). See Supplementary Materials for the distribution of the age of the participants (Figure S1) and participant information (age, gender, SES, HL onset, ML onset) of those from English-dominant vs. German-dominant environment (Table S7). The vast majority of the HSs were second-generation immigrants, and eight participants were third-generation and two participants were fourth-generation HSs. The HSs were all exposed to Japanese before the age of three (Mean = 1.11 months, SD = 4.73, Range = 0 – 36)¹. The HSs from an English-dominant environment

¹ The reason why we included those who were not exposed to the HL from birth is that there were some respondents who misinterpreted the question (“When was the child first exposed to the heritage language (Japanese)/majority language (German or English)?”) and did not enter “0” for the HL or the ML (e.g., 36 months for HL and 60 months for ML). This would mean that the child was not exposed to any language/deprived of any language from birth, which is doubtfully the case. It is likely that the respondents entered the age in which the child started **speaking** in the HL, or the age in which the child was exposed to HL **outside** of the home e.g., daycare. The misinterpretation of this question has also been reported by the Q-bex questionnaire team when they conducted assessments on validating the questionnaire. However, since HL onset was excluded from the Factor Analysis most likely due to the low variance in the responses (i.e., most people answered 0); we decided to include these participants in the subsequent regression analyses where we are

lived in the following countries: USA (n = 180), Australia (n = 46), Canada (n = 45), United Kingdom (n = 33) and those from German-dominant environment came from Germany (n = 104) and Switzerland (n = 6)². Their mean onset to the societal majority language (English or German) was 11.95 months (SD = 18.69; Range = 0 – 78 months). We only included children who attended schools in the majority language of the society. Their SES was also measured via the main caretaker's final education from a scale of 0 to 4 (Mean = 3.07, Range = 0 – 4). The HS participants were recruited via personal network, Japanese Saturday schools, Facebook groups, and Japanese communities abroad and they were compensated 10 euros/pounds/dollars for their participation.

2.2 Questionnaire

We collected extensive information about the participants' language background experience using the Quantifying Bilingual Experience (Q-bex) questionnaire (De Cat et al., 2023). Q-bex is a newly established, user-friendly online questionnaire that can be customized in various ways to facilitate administration and ensure that the desired level of detail is acquired. Q-bex comprises two mandatory modules (background information and risk factors) and, in addition, five optional ones (language exposure and usage, language proficiency, richness of linguistic experiences, attitudes, and satisfaction with the child's language, language mixing), which have sub-modules that can be individually chosen or

mainly interested in how the interaction between Age and Latent Experience Variables predicts comprehension and production of classifiers and case-marking.

² Herein, we do not address the potential for cross-linguistic influence (CLI) given differences between the two societal majority languages (e.g., effects that might obtain from differences in overt morphological case between German and English). Doing so pushes us outside the scope of the present paper's theoretical remit. The present aim is to unpack the effects of the relative contributions of variables that condition engagement with the HL, Japanese, itself. Examining this would effectively force a sub-group to sub-group aggregate comparison, and yet doing so would detract from the individual differences approach we take to properly investigate our main foci (Rothman et al., 2023). Pursuing questions of CLI is interesting, to be sure, and we will pursue such questions in a separate paper with unique research questions, distinct approaches to the analysis and proper contextualization of its theoretical relevance. That being said, we report the findings on the effects of majority language (English, German) and its interaction with latent experiential factors on grammatical competence/performance in the Supplementary Materials, given that some experiential latent factors may matter more for classifier/case-marking/passive competence in one majority language context than the other.

omitted based on the specific needs of researchers. We selected specific sub-modules that we hypothesized to be particularly relevant in the context of our study. These selected sub-modules included: current (input) estimates, cumulative (input) estimates, age and place of first exposure, overheard speech at home, proficiency (no reference group), activities, caregiver’s education, estimated diversity of speakers, preferred language, language mixing at home and outside between interlocutors.


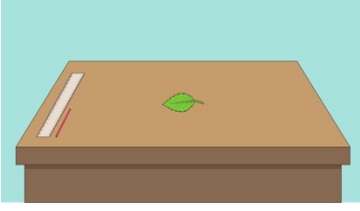


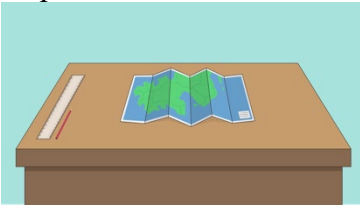

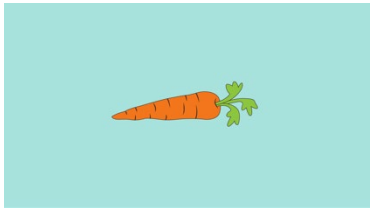
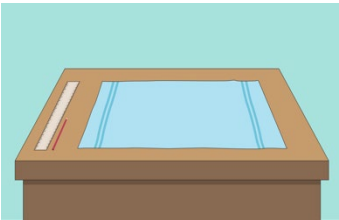

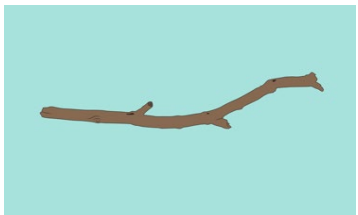
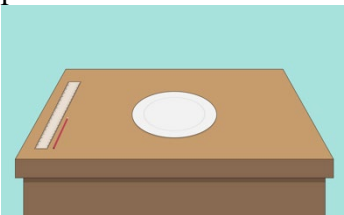


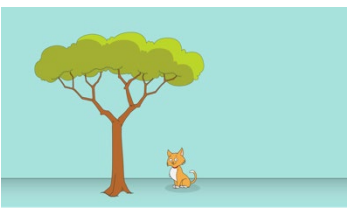
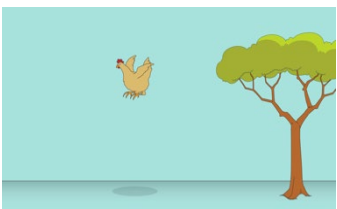
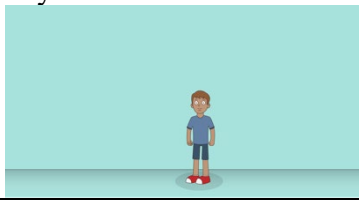
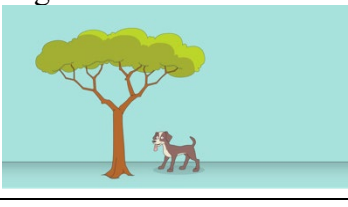
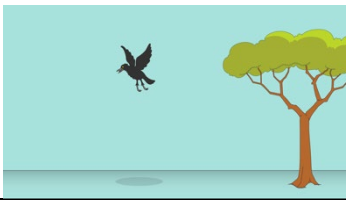
2.3 Classifier tasks



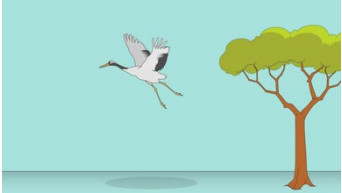


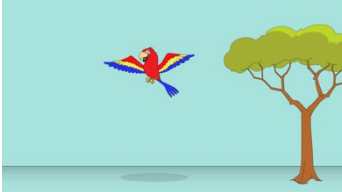
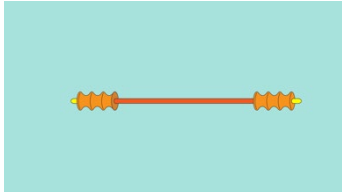


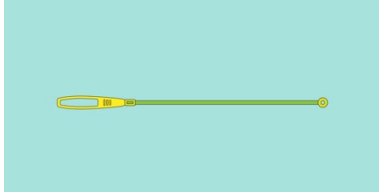
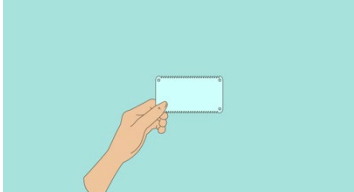
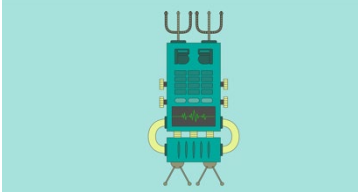
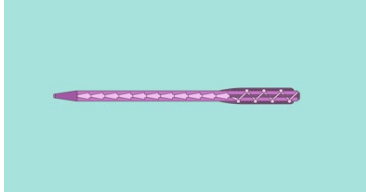


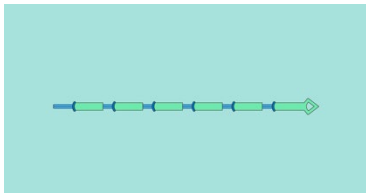
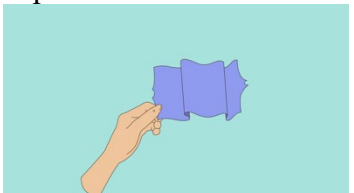
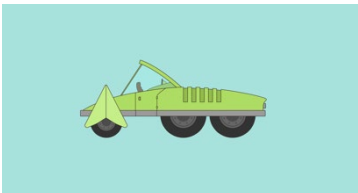
We tested the HSs’ classifier knowledge and use via a comprehension task and a production task. We tested six classifier categories: *-hon* (long, thin objects), *-mai* (flat, thin objects), *-dai* (machines), *-ri* (humans), *-hiki* (small animals), and *-wa* (birds). These classifiers were selected because they possess clearly defined and noticeable perceptual characteristics and are visually distinctive and frequently encountered in everyday life.

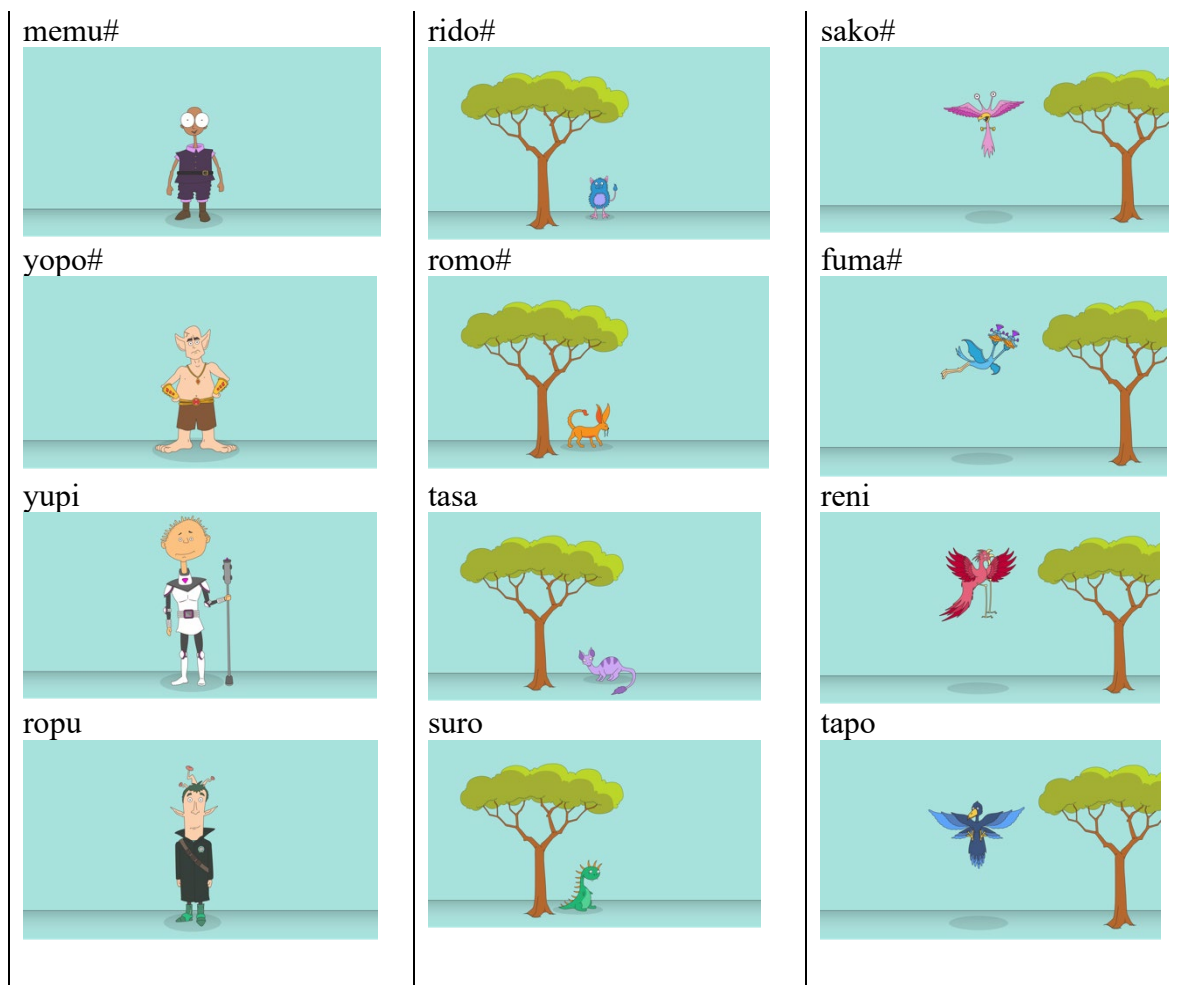
The comprehension task consisted of a total of 48 target items, divided equally into 24 familiar items and 24 nonce items. Within each category of familiar and nonce items, we included four items for each classifier category, as outlined in Table 1. We matched the frequency of familiar items within each classifier category, although it proved challenging to achieve frequency matching across categories. This is due to the fact that certain categories (e.g., *-dai* for machines and *-wa* for birds) are exclusively associated with specific items or animals that are less common in the input when compared to those belonging to more general categories like *-hon* (long, thin objects) or *-hiki* (small animals).

Table 1. Full list of classifier items in the comprehension task
= subset of items used in the production task

	Familiar items	
Inanimate Classifiers		
Hon (1D long thin)	Mai (2D flat)	Dai (machines)

<p>banana#</p> 	<p>leaf#</p> 	<p>TV#</p> 
<p>pencil#</p> 	<p>map#</p> 	<p>phone#</p> 
<p>carrot</p> 	<p>towel</p> 	<p>bicycle</p> 
<p>branch</p> 	<p>plate</p> 	<p>car#</p> 
<p>Animate Classifiers</p>		
<p>Ri (humans)</p>	<p>Hiki (animals)</p>	<p>Wa (birds)</p>
<p>girl#</p> 	<p>cat#</p> 	<p>chicken#</p> 
<p>boy#</p> 	<p>dog#</p> 	<p>crow#</p> 

<p>man</p> 	<p>mouse</p> 	<p>crane</p> 
<p>woman</p> 	<p>fish</p> 	<p>parrot</p> 
<p>Nonce items</p>		
<p>Inanimate Classifiers Hon (1D long thin) sonu#</p> 	<p>Mai (2D flat) poru#</p> 	<p>Dai (machine) naso#</p> 
<p>yapu#</p> 	<p>mupi#</p> 	<p>koni#</p> 
<p>honi</p> 	<p>nopu</p> 	<p>gemi</p> 
<p>chiza</p> 	<p>napu</p> 	<p>nefu</p> 
<p>Animate Classifiers Ri (humans)</p>	<p>Hiki (animals)</p>	<p>Wa (birds)</p>



The labels assigned to the nonce items underwent a norming process involving 46 adult native Japanese speakers. They rated the level of meaning conveyed by the presented words on a scale ranging from 1 to 4, where 1 indicated no discernible meaning, and 4 indicated a clear meaning for the word. All nonce labels utilized in Table 1 received ratings below 1.4.

Regarding the images used for both familiar and nonce items, a norming procedure was carried out twice, initially involving 49 adult native Japanese speakers and 59 participants in the second round of piloting. In both rounds, the participants were instructed to provide the appropriate classifier for each image. In the initial round, some nonce images exhibited relatively low classifier agreement, ranging from 8% to 100%, while all familiar images consistently elicited 80% agreement or higher. Consequently, we excluded images

that generated less than 70% agreement and introduced new nonce images. These newly introduced stimuli, along with the retained ones from the first round, underwent a reevaluation by another set of native Japanese raters. Images that achieved a consensus of over 70% agreement in the second round were selected for inclusion in the experiment, as in Table 1.

2.3.1 Classifier comprehension task

In the comprehension task, participants were asked to select the correct picture that matched the target classifier from two available options. Each trial followed a specific sequence: the participant initiated the audio (consisting of a numeral followed by a classifier) by clicking on a small alien icon located at the bottom of the screen. Subsequently, they maneuvered the cursor to click on either the picture in the top left or the one in the top right of the screen. For example, as shown in Figure 2, when the participant heard "*ichi-mai* (one flat-CL)" after clicking on the alien, they had to move their cursor to the right to select the corresponding target picture (in this case, the plate).

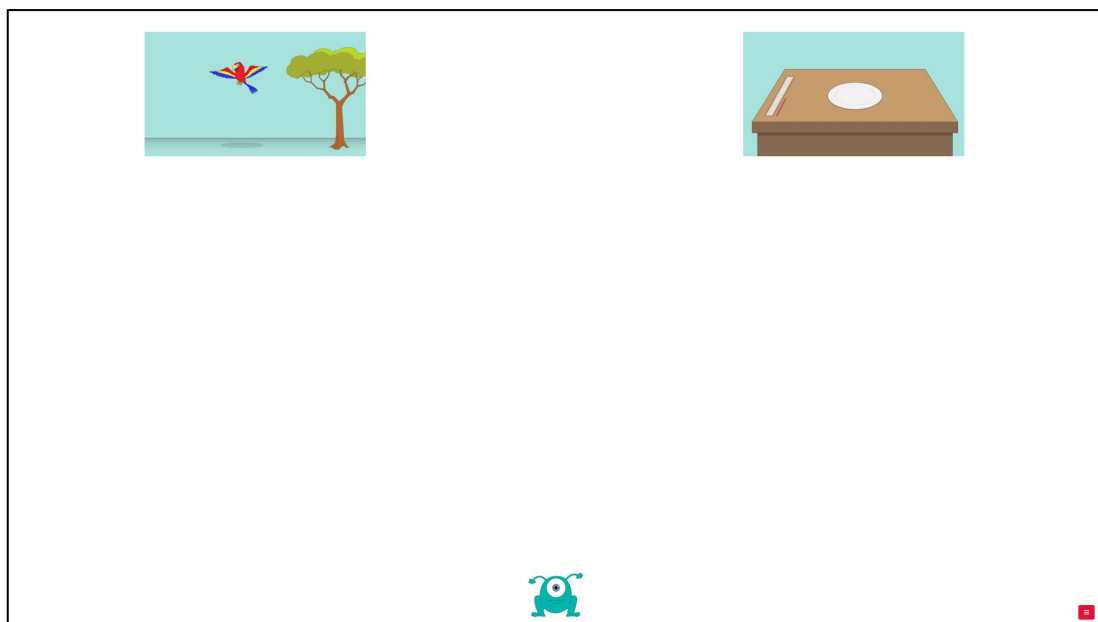


Figure 2. Illustration of the classifier comprehension task

There were 48 classifier items (24 familiar and 24 nonce items) as shown in Table 1 and 32 case-marking items (as described below in Section 2.4.1). The target and competitor items were either exclusively familiar or exclusively nonce items. We deliberately avoided including combinations such as familiar-target & nonce-competitor (or vice versa) pairs since children could tend to favor the familiar item when presented alongside a nonce item, irrespective of the classifier used.

Additionally, we took steps to balance the animacy of the target-competitor pairs, ensuring an even distribution across trials. Specifically, one-fourth of all trials fell into each of the following categories: (a) target-animate & competitor-inanimate (mismatched pairs), (b) target-inanimate & competitor-animate (mismatched pairs), (c) target-animate & competitor-animate (matched pairs), and (d) target-inanimate & competitor-inanimate (matched pairs). This manipulation was based on findings by Yamamoto and Keil (2000), which indicated that Japanese children performed differently when the animacy of the target-competitor pairs was either matched or mismatched.

2.3.2 Classifier production task

In the production task, the participants were given specific instructions to count the number of items shown in a picture. The items in question always followed a sequence of one, two, and three. First, the participants watched a video in which a researcher explained the task in Japanese. They also underwent a microphone check to ensure the recording system was functioning correctly. Subsequently, the participants completed two practice trials using classifiers that were different from the target classifiers (specifically, *-too* for large animals and *-hai* for a glass/cup of liquid). In each trial, the participants heard the name of an item (e.g., *kore wa yopo desu* “This is a yopo”). They were then instructed to count the items displayed on the screen, one by one, as illustrated in Figure 3. After two classifier trials, they

were asked to describe two sets of pictures in which we also tested their knowledge of case-marking/passive structures (as detailed below in Section 2.4.2).

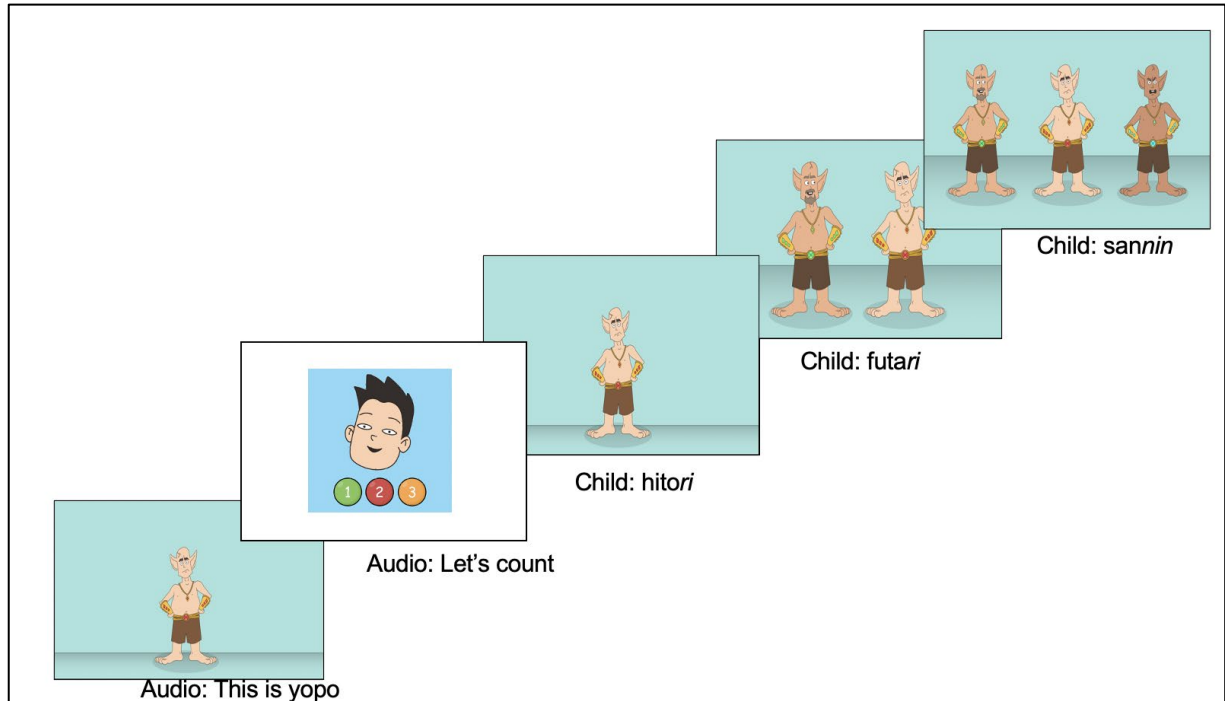


Figure 3. Illustration of the classifier production task

In the production task, we used half of the items from the comprehension task ($N = 24$, consisting of 12 familiar items and 12 nonce items), as indicated in Table 1 with a hash key to reduce the duration of the experiment and minimize potential participant fatigue. The presentation of trials was randomized across participants. The audio recording began automatically when the picture was displayed to the participants and stopped when the participant clicked the button to proceed to the next trial.

2.3.2.1 Transcription and coding

Four research assistants who are native speakers of Japanese and received training in linguistics transcribed and coded the data. 10% of the data was coded for reliability purposes and the inter-rater reliability among the four research assistants was 97.19%. A final quality

control check was conducted by one of the principal investigators (PI), who is also a native Japanese speaker.

We coded the data using a binary choice (0,1) based 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 used the correct classifier. As detailed further in the procedures section below, given the age of the participants, parental supervision was required for those under 12. Although parents were instructed not to interfere with the automated experiments, any interferences that the parents made were automatically recorded for each trial. Instances where a parent provided the classifier to the child, such as by interjecting and asking the child, "How many CL?" were excluded from the analysis. In sum, they accounted for only 0.3% of the data. Additionally, 4.3% of the data that were either unintelligible or lacked audio content were excluded from the analysis.

2.4 Passive and case-marking tasks

2.4.1. Case-marking/passive comprehension task

We embedded the comprehension task of case-marking/passive in the same paradigm as described above in the comprehension task of classifiers. This way, case-marking/passive items serve as fillers for classifier items and vice-versa, allowing us to test two target structures in one task, hence minimizing the length of testing. We tested comprehension of case-marking/passive structures by having both canonical (SOV) and non-canonical (OSV) word order for active and passive sentences as outlined in Table 2 below.

Table 2. Conditions, example sentences, and English translations for the stimuli used in comprehension task

	Condition	Example Sentences	English translation
(1)	Active	inu (ga) niwatori (o) oshi-ta	Dog pushed the chicken
	Canonical	dog (NOM) chicken (ACC) push-PAST	

(2)	Active	niwatori (o) inu (ga) oshi-ta	Dog pushed the chicken
	Non-canonical	horse (ACC) gorilla (NOM) push-PAST	
(3)	Passive	niwatori (ga) inu (ni) o-sare-ta.	Chicken is being pushed by the dog
	Canonical	chicken (NOM) dog (DAT) push-PASS-PAST	
(4)	Passive	inu (ni) niwatori (ga) o-sare-ta.	Chicken is being pushed by the dog
	Non-canonical	dog (DAT) chicken (NOM) push-PASS-PAST	

Each condition included four items with a total of 16 sentences for target conditions and 16 other sentences with dative structures in order to balance the item numbers between classifier (48 sentences) and case-marking and passive sentences (32 sentences), totaling 80 sentences. As presented in Figure 4, and in a similar vein to the classifier comprehension task, two pictures were presented side by side on a screen. This is accompanied by a spoken utterance such as in condition (2): “niwatori o-ACC inu ga-NOM oshita.” If the participants interpret the case morphology (rather than relying on canonical word order), then they should choose the picture on the right-hand side of the screen with the dog pushing the chicken. In a similar vein, if the participants hear the passive structure such as condition (3): “niwatori ga-NOM inu ni-DAT o-sare-ta-PASS-PAST”, then participants should choose a picture with a chicken being pushed by the dog (i.e., dog pushing the chicken). If the participant instead chooses the picture with the chicken pushing the dog, then this indicates that they do not have the representation or process in the moment the *rare* morphology as passive voice, thus interpreting the utterance as an active sentence.

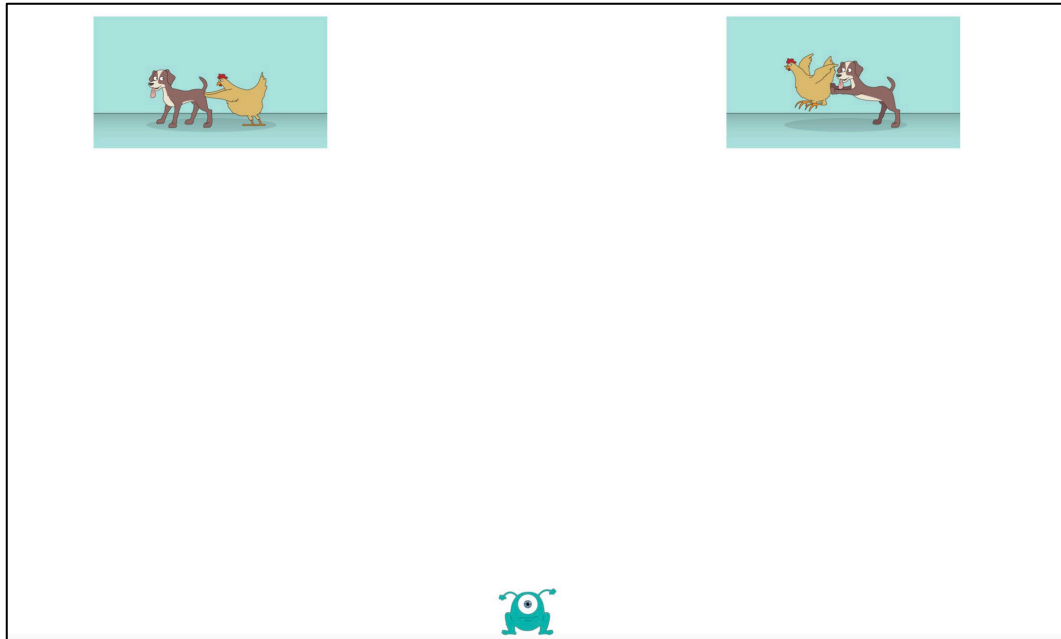


Figure 4. Illustration of the case-marking comprehension task

For both the classifier and case-marking/passive comprehension tasks, there was no predefined time limit for each trial, and participants automatically proceeded to the next trial upon clicking on a picture. They were instructed to select the picture as accurately and quickly as possible. We also counterbalanced the placement of the target picture/item (either on the right or left upper screen), and the order of trials was randomized across participants. The comprehension task began with a comprehensive instructional video explaining the task, followed by two practice trials (one that did not include the target classifier and another that involved simple sentence comprehension e.g., the cat looked up). There was a short break after half of the trials were completed. We analyzed participants' accuracy in selecting the target picture for subsequent analyses. Although reaction time (RT) of the comprehension task was also recorded, we opted for reporting the RT descriptive results in the Supplementary Materials (see Figure S2 and S3).

2.4.2. Passive production task

Similarly to the comprehension task, the production task for case-marking/passives was embedded within the production task of the classifiers, so that each served as a filler for the other. There were 24 different sets of pictures, consisting of 12 pictures that were aimed to elicit an active structure (with an agent focused question), and 12 pictures were aimed to increase the probability of eliciting a passive structure (with a patient focused question), as in Figure 5. Each picture was shown to the participant one by one on a computer screen. In order to make the production experiment coherent across the classifier and case-marking/passive tasks, we included both nonce and familiar animate characters that were introduced in the two sets of classifier trials prior to the case-marking/passive trials. We included audio buttons on the bottom right corner of the screen (explained/demonstrated to participants in the instructional video), providing the name of the familiar and nonce characters (“crow” and “koni” in the example in Figure 5) in case the participants could not recall the names of particular characters. We also made sure that the target character (i.e., agent in the agent-focused questions and patient in the patient-focused questions) was always located on the left-hand side of the picture as well as for the buttons to probe for agent- and patient-focused answers.

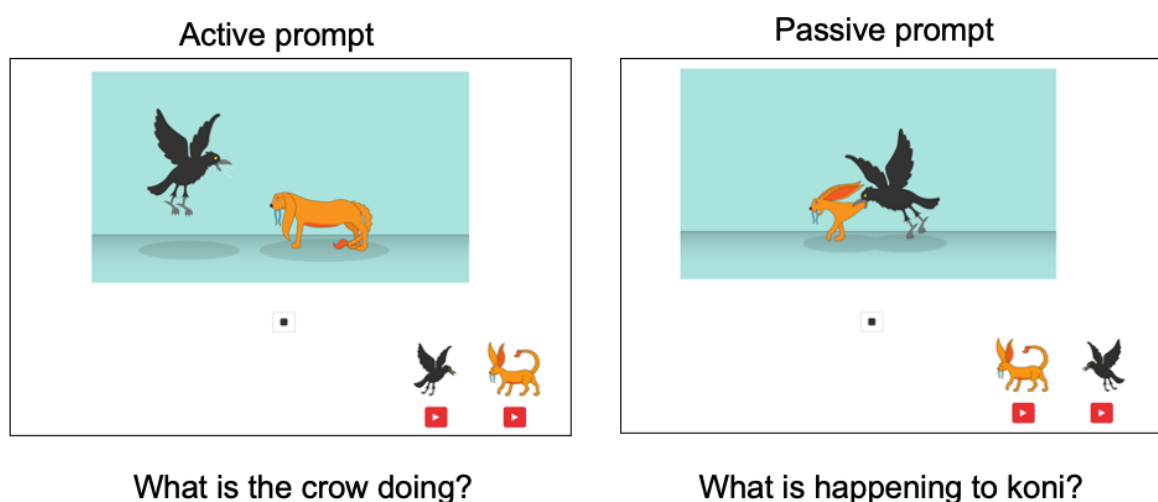


Figure 5. Illustration of the case-marking/passive production task

2.4.2.1 Transcription and coding

All utterances were transcribed and coded by four native speakers of Japanese and checked over by the native Japanese PI. The inter-rater reliability among the four research assistants (based on 10% of the coded data) was 95.72%. Any structure with the passive *-rare* morphology on the verb was coded as a use of a passive structure, regardless of phonological errors (e.g., “-name *rara* teiru” instead of “-name *rare* teiru”) or incorrect combination of verb and *-rare* morphology (e.g., “-kikku *sare* teiru” instead of “-ke *rare* teiru”). Otherwise, they were coded as an active structure. We also coded for the accuracy of the responses, namely, whether the description of the picture was correct and comprehensible to the listener. For instance, utterances such as “*neko kawaii* (cat is cute)” or “*koyatte shiteiru* (doing it like this)” were coded as incorrect responses since they are not relevant answers to the questions asked. In a similar vein to the classifier coding scheme, utterances in which a parent provided the passive *-rare* morphology to the child, such as by asking the child, “*Nani sare te iruno?*” (What does-PASS? What is being done?) were excluded from the analysis. In sum, they accounted for only 0.9% of the data. Additionally, 7.74% of the data that were either unintelligible or lacked audio content were also excluded from the analysis.

2.5 Procedure

Both tasks were designed and executed using the Gorilla (<https://gorilla.sc/>) platform and participants engaged in the study from the comfort of their own homes. We ensured that participants could only access the experiment from their laptops or computers and not from mobile phones or iPads. The experiment began with viewing a general introduction video, which instructed them to be in a quiet environment free from distractions. Parents were explicitly instructed not to provide their children with answers, and we requested them to supervise their children if they were under the age of 12. After obtaining consent from the parents, the children were presented with a short animation that included a cover story video.

This cover story involved two astronauts, Ken and Lisa, who had their spaceship stolen by aliens. The children were tasked with assisting Ken and Lisa in reclaiming their spaceship by completing various missions that featured different creatures. It's important to note that we always administered production tasks before comprehension tasks. This decision was made because the comprehension task revealed the target classifiers, and we aimed to prevent any learning/priming effects from the comprehension task from influencing the children's performance in the production task. Upon the completion of both the production and comprehension tasks, parents were asked to complete Q-bex and a compensation form. The entire online experiment can be accessed through the Gorilla open materials page, accessible via the provided link: <https://app.gorilla.sc/openmaterials/686845>

3. Results

3.1 Factor analysis on the questionnaire

We first ran a factor analysis to extract underlying latent factors from the Q-bex questionnaire and to use the factor scores to predict classifier and case-marking comprehension and production in the subsequent analysis. We included questions in the Q-bex data that were related to the heritage language (Japanese) and not their societal language (English or German). These questions are listed in the Supplementary Materials Table S1.

As a first step in running an EFA, we centered and scaled all responses and ran a Kaiser-Meyer-Olkin (KMO) Test (a measure of how suited the data is for factor analysis), which showed that only one variable (HL_onset) was below the value of 0.6, and thus this item was omitted from further analysis. Thirty-one items were analyzed with an ordinary-least-squares minimum residual approach to EFA using an oblique rotation (promax), allowing for factors to correlate. The eigenvalue method (i.e., Kaiser's rule) suggests three factors to be extracted, while the scree plot suggested around six to seven, and the parallel plot six. Thus, we determined to obtain six factors in the subsequent analysis (see

Supplementary Materials Table S2 for the factor loadings after rotation for the final analysis). Inspection of the clustering items suggests that Factor 1 represents “Community” (how much HL exposure and use the child has with friends and adults in the community), Factor 2 represents “School” (how much HL exposure and use the child has with friends and teachers at school), Factor 3 represents “Immersion” (how much HL exposure and use the child has with other children and adults during holiday trips), Factor 4 represents “Proficiency” (self-rated HL proficiency in speaking, understanding, reading, and writing), Factor 5 represents “Literacy” (frequency in HL reading, homework, and school lessons) and Factor 6 represents “Home” (how much HL exposure and use the child has with the main caretaker at home and the final education of the main caretaker in the HL). These six factors accounted for 59.30% of the total variance. We then extracted the factor scores from the factor analysis and used them as predictors for the subsequent analyses with comprehension and production tasks for all grammatical domains.

3.2 Classifier results

3.2.1 Statistical analyses

We examine whether: (a) there is a difference in the developmental trajectories of comprehension and production of familiar and nonce items, (b) latent experiential factors predict the development of classifier comprehension and production, (c) latent experiential factors predict the development of classifier comprehension and production differently for familiar and nonce items. Thus, we constructed a generalized linear mixed effects (glmer) model with accuracy as a binary dependent factor and Familiarity (familiar, nonce), Age, Community, School, Immersion, Proficiency, Literacy, Home, as well as a two-way interaction between Age and all other latent experiential factors, as well as a three-way interaction between Age, Familiarity, and all other latent experiential factors for both comprehension and production analyses. All numerical fixed factors were centered. We also

included random intercepts for item and participant and a familiarity slope for participant intercept. This resulted in the following model syntax (for both comprehension and production): `glmer (accuracy ~ age*familiarity*(community + school + immersion + proficiency + literacy + home) + (familiarity|participant) + (1|item)`. See Supplementary Materials for the effect of the majority language (English, German) on classifier comprehension (Table S8) and production (Table S9) data.

3.2.2 Classifier comprehension

We first present the descriptive results of the accuracy on the comprehension task as indicated in Figure 6 (See Supplementary Materials Figure S2 for reaction time results). The overall accuracy is high, with all classifier types eliciting accuracy of at least 89% accuracy for both familiar and nonce items. There were no significant differences between familiar and nonce items ($B = .96, p = .85$) for classifier comprehension accuracy.

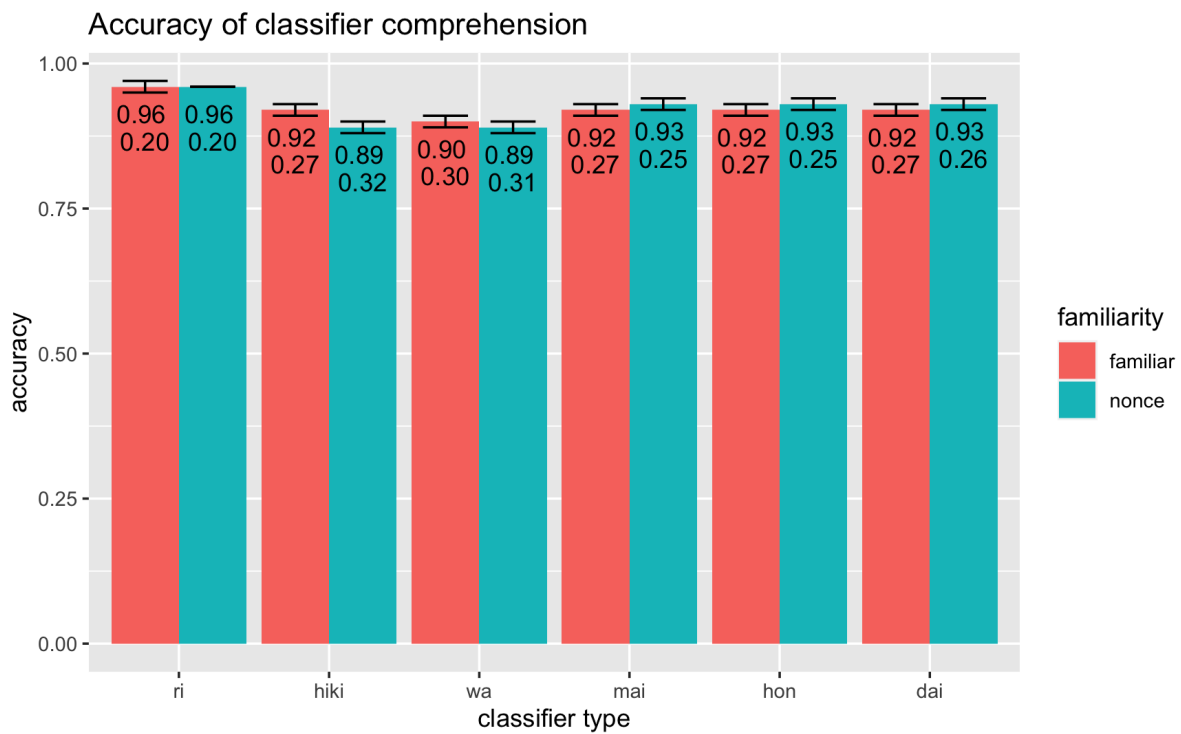


Figure 6. Accuracy (top) and standard deviation (bottom) of classifier comprehension

The model summary can be found in Supplementary Materials Table S3. There were no significant two-way or three-way interactions and only a significant main effect of Proficiency ($B = 1.74, p = .001$), Immersion ($B = 1.55, p = .005$), and Literacy ($B = 1.58, p = .001$). These estimates are all positive, indicating that comprehension accuracy of classifiers improves with age and the higher the proficiency and the more immersion experiences and literacy practices HSs have in Japanese, the better their performance is on classifier comprehension (irrespective of its familiar or nonce status).

3.2.3 Classifier production

The aggregate accuracy of target classifier production is presented in Figure 7. In contrast to the comprehension results, there is a wide range of production accuracy across various classifier types, with *-ri* (human) eliciting the highest accuracy followed by *-hiki* (small animals), *-hon* (long, thin objects), *-wa* (birds), *-mai* (flat, thin objects), and *-dai* (machines) respectively. Descriptively, familiar items had higher accuracy than nonce items (except for the *-dai* classifier), however statistically, there were no significant differences between familiar and nonce items ($B = .74, p = .40$). The model summary of the classifier production can be found in Supplementary Materials Table S4.

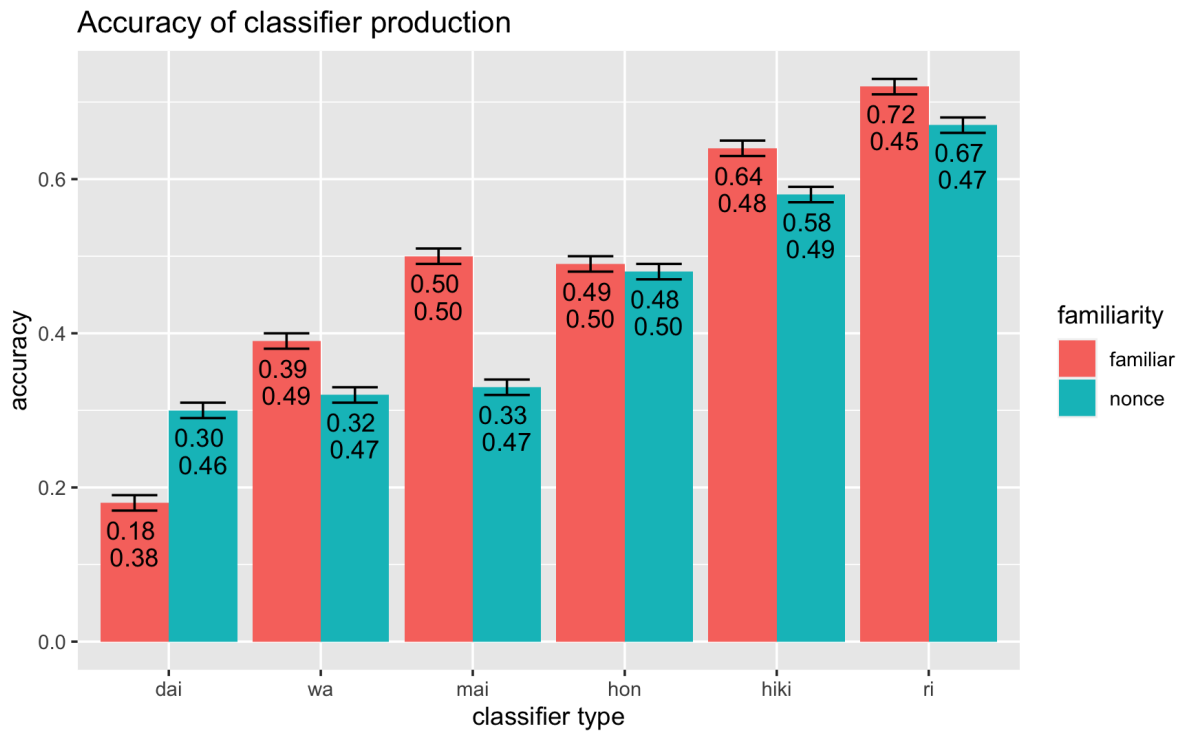


Figure 7. Accuracy (on top) and standard deviation (on bottom) of classifier production

First, there was a significant interaction between Age and Familiarity ($B = .93, p = .003$) as illustrated in Figure 8. The S-curve suggests that younger children perform at floor regardless of whether an item was familiar or nonce. However, as children transition from middle childhood through adolescence (from around age seven to eighteen), familiar items exhibit a more pronounced developmental progression compared to nonce items. That is, producing classifier-noun pairings that are available in their input is easier than extending the classifier meaning to novice items in production.

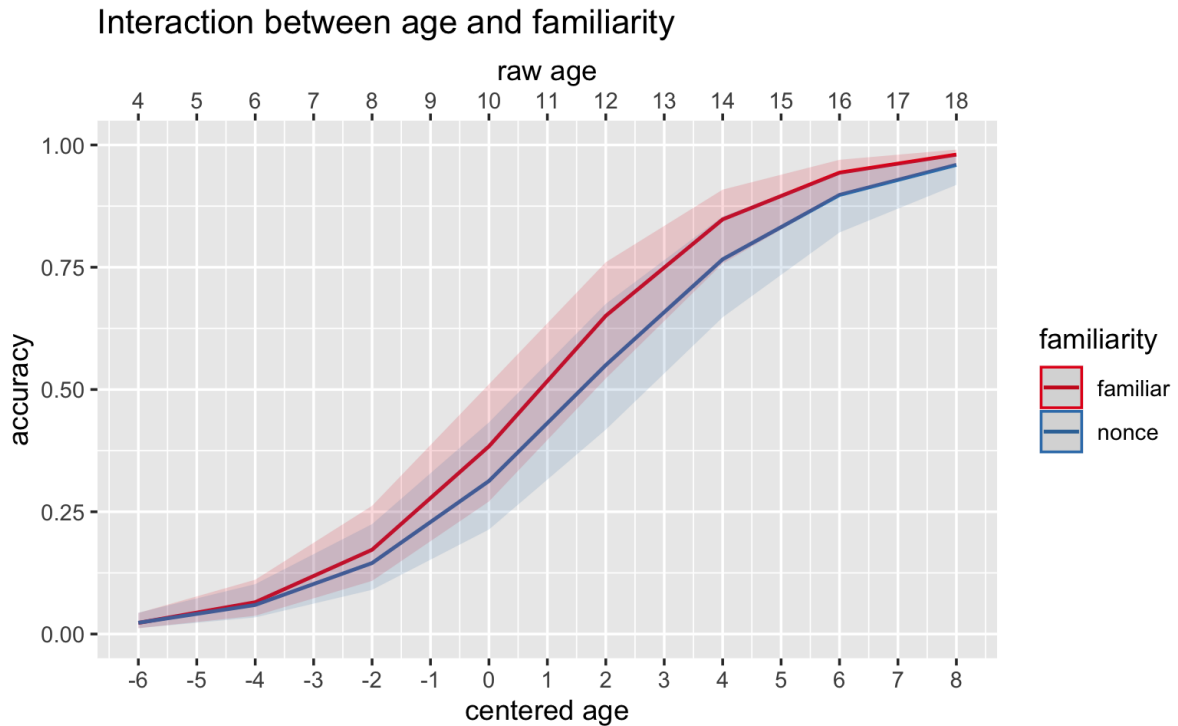


Figure 8. Two-way interaction between Age and Familiarity (familiar, nonce) on classifier production accuracy. X-axis on the top indicates raw age and the x-axis on the bottom indicates centered age.

In terms of the effects of experiential factors, there was a main effect of Proficiency ($B = 2.40, p < .001$), Immersion ($B = 1.50, p = .01$), Literacy ($B = 2.10, p < .001$), and as well as a significant two-way interaction between Age and Community ($B = 1.15, p = .006$) and a three-way interaction between Age, Immersion, and Familiarity ($B = .95, p = .04$). As for the effect of HL engagement in the community/society (Figure 9), we see that those with *less* HL engagement in the community display higher accuracy in classifier production until around age eleven. However, from age eleven and onwards, we see a flipped effect in which HSs with *more* HL engagement in the community exhibit higher classifier production accuracy than those with less HL engagement. As for the effect of Immersion on the productive development of classifiers, children who experienced richer immersion experiences (trips in Japan) have higher accuracy in classifier production up until around 16 years old, but this effect dramatically wanes off thereafter. Moreover, immersion experience seems to have a

greater effect on the development of production accuracy in nonce items than familiar items (i.e., the difference between the red, blue, and green lines in Figure 10, which indicate the degree of immersion experiences is larger for nonce than familiar items).

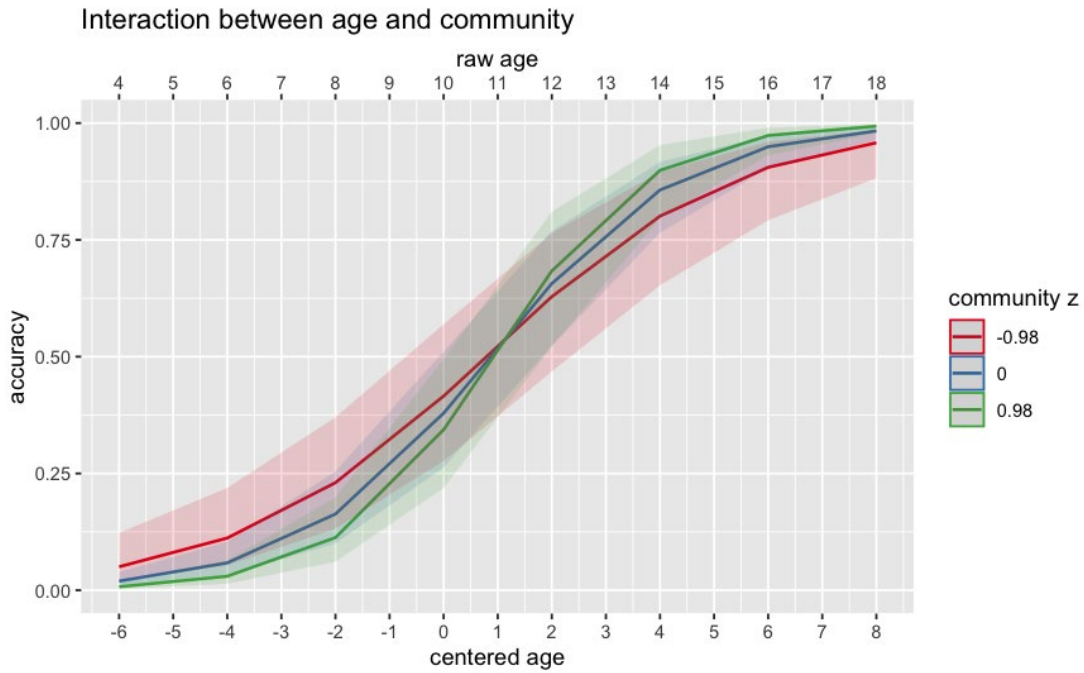


Figure 9. Two-way interaction between Age, Community (centered) on classifier production accuracy. X-axis on the top indicates raw age and the x-axis on the bottom indicates centered age.

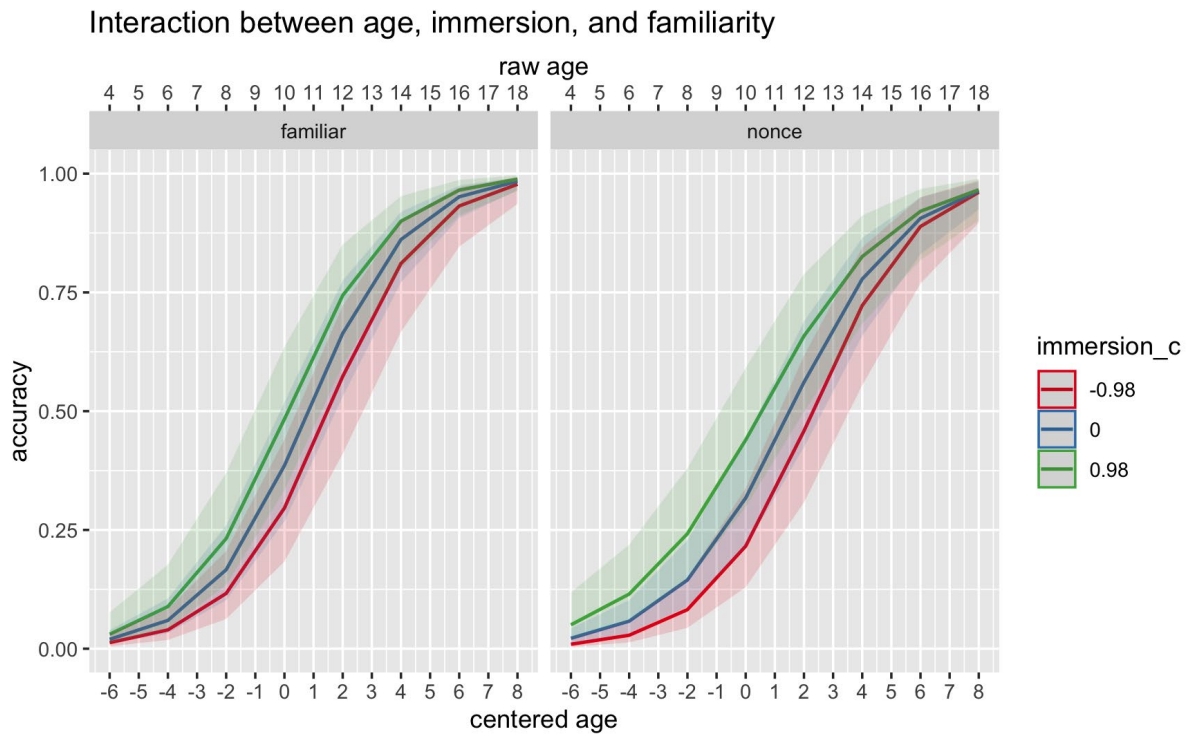


Figure 10. Three-way interaction between Age (centered) Immersion (centered) and Familiarity (familiar, nonce) on classifier production accuracy. X-axis on the top indicates raw age and the x-axis on the bottom indicates centered age.

3.3 Case-marking/passive results

3.3.1 Statistical analyses

We examine whether: (a) there is a difference in the developmental trajectories of canonical vs. non-canonical and active vs. passive items, (b) latent experiential factors predict the development of case-marking structures, (c) latent experiential factors predict the development of case-marking structures differently for canonical vs. non-canonical and active vs. passive items. To this end, we built a glmer model with accuracy as a binary dependent factor and Voice (active, passive), Canonicity (canonical, non-canonical), Age, Community, School, Immersion, Proficiency, Literacy, Home, as well as a two-way interaction between Age/Voice/Canonicity and all other latent experiential factors and a three-way interaction between Age, Voice/Canonicity, and all other latent experiential factors. All numerical fixed factors were centered. We also included random intercepts for item and participant and by-participant slopes for Canonicity and Voice. This resulted in the

following model syntax (for comprehension): $\text{glmer}(\text{accuracy} \sim \text{age} * \text{voice} * (\text{community} + \text{school} + \text{immersion} + \text{proficiency} + \text{literacy} + \text{home}) + \text{age} * \text{canonicity} * (\text{community} + \text{school} + \text{immersion} + \text{proficiency} + \text{literacy} + \text{home}) + (\text{caononicity} + \text{voice} | \text{participant}) + (1 | \text{item}))$.

Regarding the production data, we examined how the experiential latent factors modulate the development of their productivity of passive structures. We constructed a glmer model with target response (i.e., whether they produced a passive structure when a patient-focused question was provided) as a dependent variable and Age, Community, School, Immersion, Proficiency, Literacy, Home, as well as a two-way interaction between Age and all other latent experiential factors as fixed effects. This resulted in the following model syntax (for production): $\text{glmer}(\text{accuracy} \sim \text{age} * (\text{community} + \text{school} + \text{immersion} + \text{proficiency} + \text{literacy} + \text{home}) + (1 | \text{participant}) + (1 | \text{item}))$. The full model summary of the comprehension and production models can be found in the Supplementary Materials Table S5 (comprehension) and S6 (production). See Supplementary Materials for the effect of the majority language on case-marking/passive comprehension (Table S10) and production (Table S11) data.

3.3.2 Case-marking/passive comprehension results

The descriptive results of the accuracy of the comprehension data are presented in Figure 11 (See Supplementary Materials Figure S3 for reaction time results). There is a main effect of Voice and Canonicity, in which active elicits higher accuracy than passive ($B = .31, p < .001$) and canonical elicits higher accuracy than non-canonical ($B = .16, p < .001$) conditions. There was also an interaction between Voice and Canonicity, and post-hoc comparisons using Tukey correction revealed that active canonical has the highest accuracy, followed by passive canonical, active non-canonical, and passive non-canonical respectively.

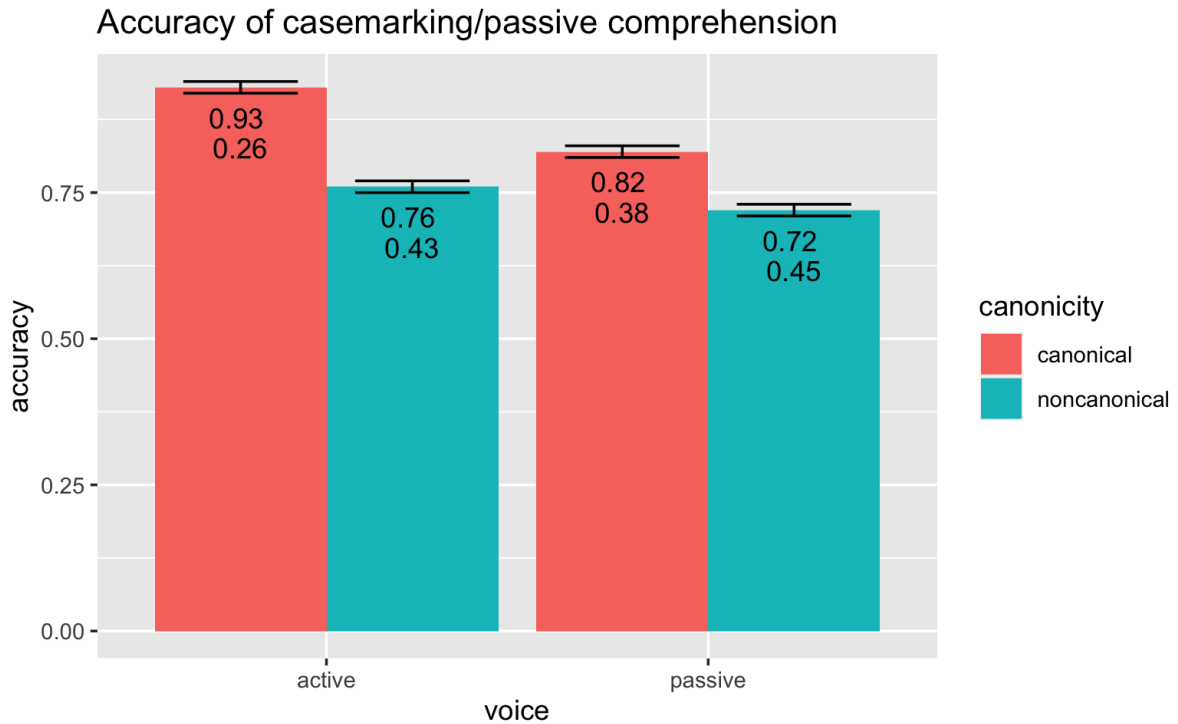


Figure 11. Accuracy (on top) and standard deviation (on bottom) of comprehension split by Voice (active, passive) and Canonicity (canonical, non-canonical)

There was a main effect of Literacy ($B = 1.35, p = .04$), and Proficiency ($B = 1.47, p = .03$) indicating that more literacy engagement and higher proficiency in the HL predicts better performance overall (regardless of Voice or Canonicity). A significant three-way interaction between Age, Voice, and Canonicity ($B = .83, p < .001$), as in Figure 12, illustrates that while active canonical, passive canonical, and passive non-canonical conditions score around chance level (50-60%) in the youngest cohort and gradually develop until they reach over 90% accuracy in the oldest cohort, the active canonical condition already elicits around 85% accuracy in the youngest cohort, leaving not too much more room for improvement as they get older.

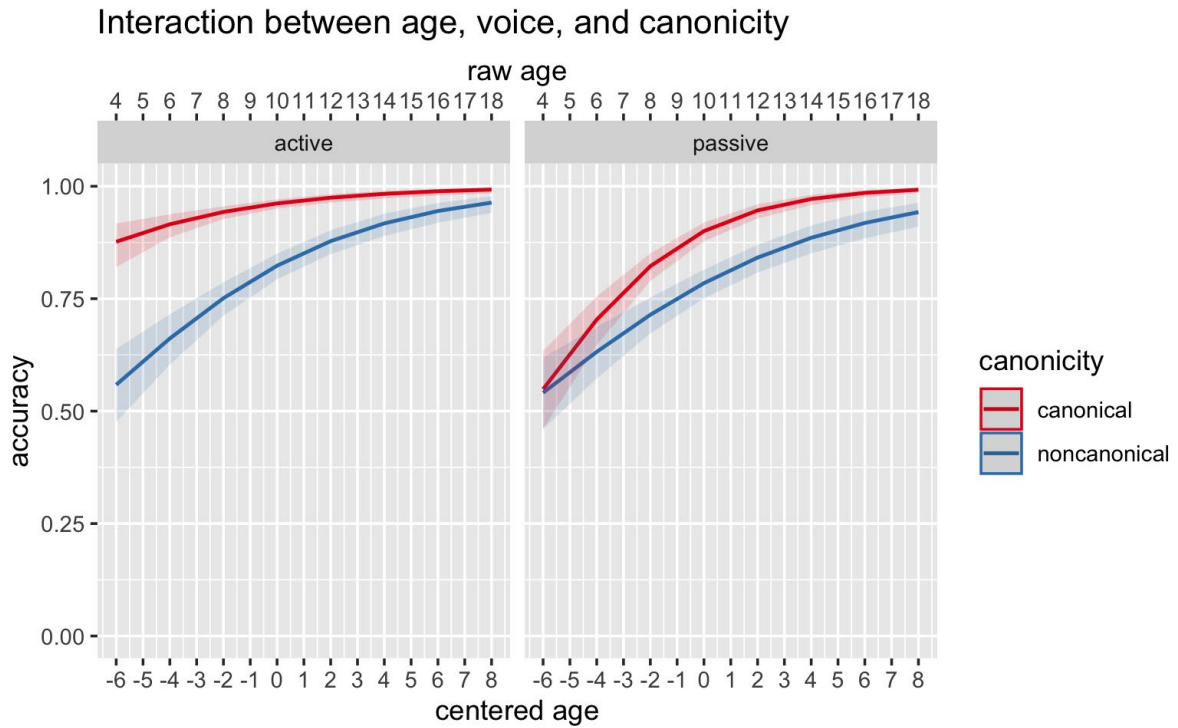


Figure 12. Three-way interaction between Age, Voice (active, passive), Canonicity (canonical, non-canonical) on comprehension accuracy. X-axis on the top indicates raw age and the x-axis on the bottom indicates centered age.

In addition, the two-way interaction between Age and Home ($B = .92, p = .02$) in Figure 13 shows that more HL engagement at home contributes to better comprehension performance until around age eight to nine. However, this effect gets weaker starting in middle childhood.

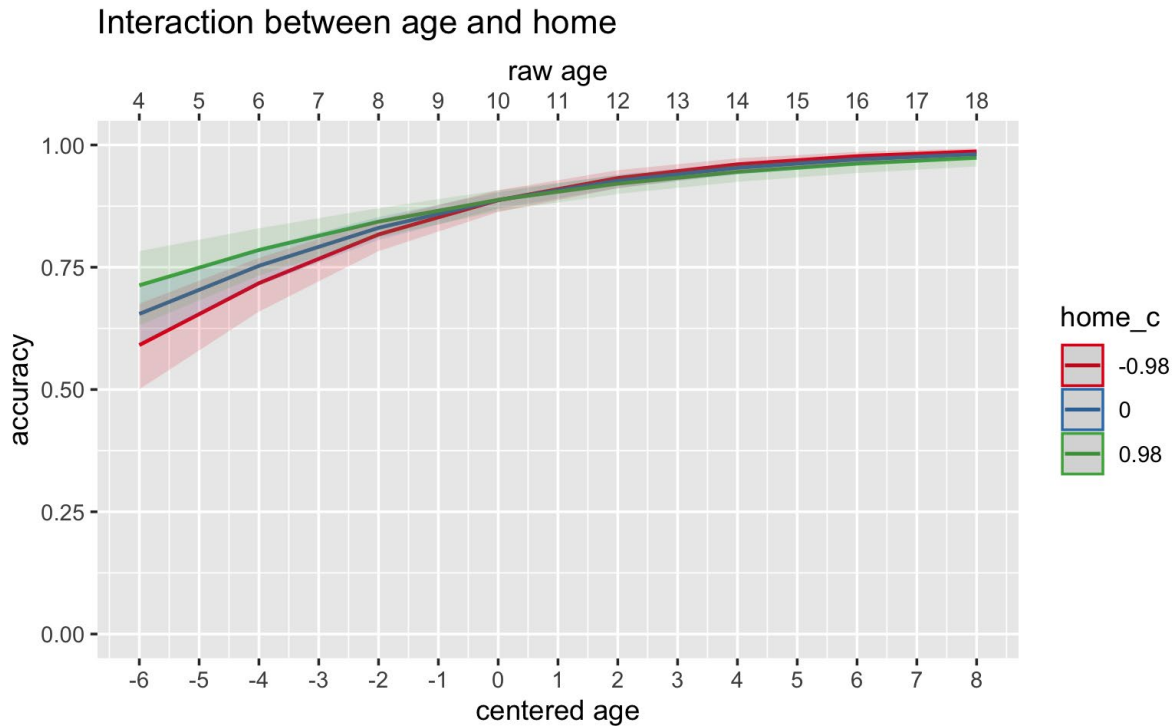


Figure 13. Two-way interaction between Age (centered) and Home (centered) on comprehension accuracy

3.3.3 Case-marking/passive production results

The HSs provided an active sentence when given an active prompt (e.g., what is the crow doing?) 98.8% (SD = 11.0) of the time; that is only 1.2% of the responses with an active prompt were given with a passive structure. In contrast, use of passive rose to 34.4% when the question had a passive prompt (a patient focused question); the remaining responses providing an active structure (65.6%). In addition, 36.3% of the participants did not produce any passive structures at all. There was a significant main effect of Immersion ($B = 1.73, p = .004$) and Proficiency ($B = 3.15, p < .001$), suggesting that richer immersion experience and higher proficiency in the HL predicted the production of passive structures. A two-way interaction was found between Age and Community ($B = 1.13, p = .03$) as illustrated in Figure 14. In a similar vein to the interaction effects shown in classifier production (Figure 9), up to the age of eleven to twelve, HSs who are less engaged with their HL in the community/society tend to produce more passive structures when provided with a passive

prompt. However, past this point a reverse trend emerges, where HSs who have more HL engagement in the community demonstrate greater production of passive structures compared to those with lower HL engagement.

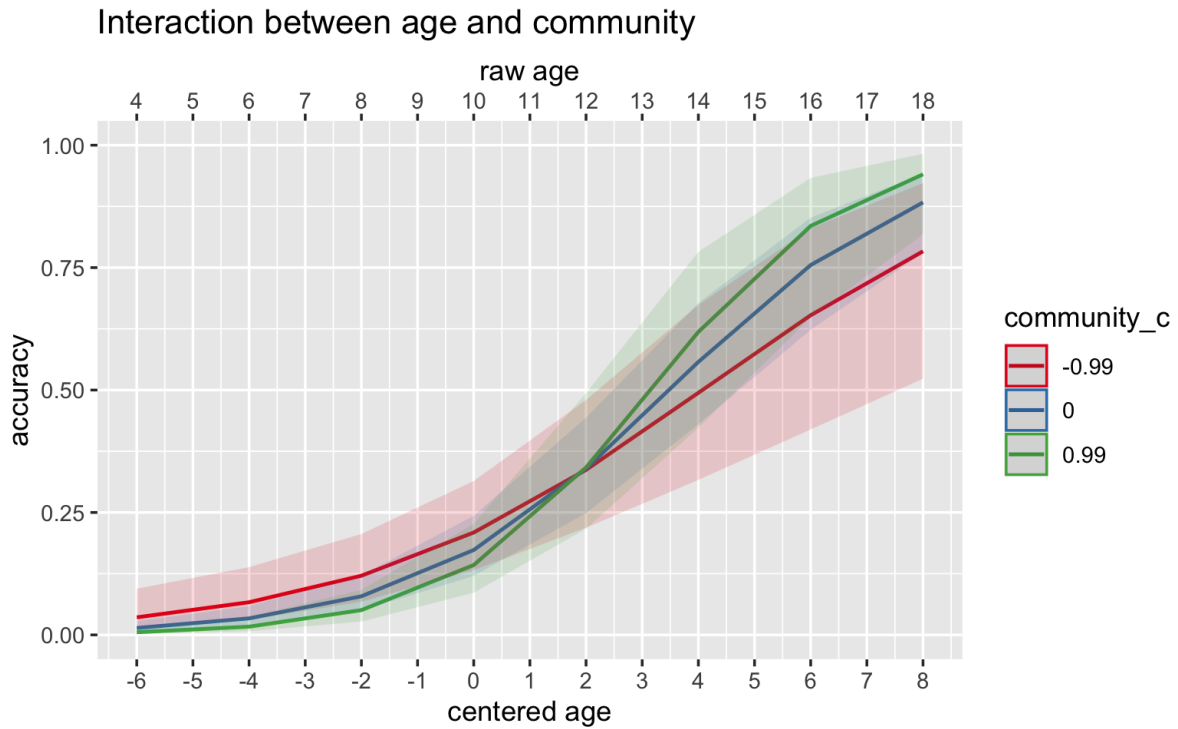


Figure 14. Interaction between Age (centered) and Community (centered) on comprehension production

4. Discussion

The current study examined what experiential factors predict the development of comprehension and production of classifiers and case-marking/passive structures in Japanese child HSs and whether they modulate grammatical competence/performance differentially from childhood through adolescence.

In terms of **classifier comprehension**, the HSs overall performed at ceiling with more than 89% accuracy on all classifier types and there were no significant differences in accuracy between nonce and familiar items. This indicates that on an aggregate level, Japanese HSs have acquired the underlying semantics of common classifiers, given that they

are able to extend the meaning of classifiers to items that they have never encountered in the input. Moreover, we found no significant interaction between age and familiarity (familiar, nonce), suggesting that the semantics of classifiers is robustly represented from the earliest ages tested (4.02 years old). This mirrors the results from Kubota et al. (2024), which tested classifier comprehension and production of Japanese monolingual children (ages 3 to 11) using the same paradigm with nonce and familiar items as in the current study. They found that Japanese monolingual children as young as three years old showed ceiling performance on both familiar and nonce items in the comprehension task.

Despite the high accuracy in classifier comprehension, HSs showed more variability (large SDs) and vulnerability (accuracy) in the **classifier production** task, in which only the familiar items for the *-ri* classifier reached an accuracy comparable to the adult monolinguals from the norming phase (70%). Although no significant differences in production accuracy were found between familiar and nonce items, there was indeed a significant interaction between age and familiarity, suggesting that the development of the semantic system of common classifiers is protracted in production when compared to comprehension. Again, this is not surprising, given that the same results are found for Japanese monolingual children (Kubota et al., 2024). Kubota and colleagues attributed these effects of asymmetry-by-modality to a processing cost that is not completely overcome until a relatively later stage in child development. This accessing cost appears to be more prominent in demanding activities such as in production, in which more cognitive resources are required to activate the grammatical and semantic information of classifiers.

Most importantly, we found Proficiency, Immersion, and Literacy to predict **classifier comprehension and production** accuracy, demonstrating that higher proficiency, richer immersion experience, and more literacy activities in the HL is crucial for the acquisition of the morphosyntax and semantics of classifiers across the age span from 4 to 18. The effect of

Community on classifier production is further modulated by age, and visual inspection of this interaction (Figure 9) indicates that children with more HL engagement in the community have lower accuracy than those with less in young childhood, but once they reach puberty (around eleven to twelve years old), this effect flips—HSs with more HL engagement in the community undergo rapid development and eventually perform better than those with less HL engagement in the community. It is interesting that the turning point for this inverted effect takes place around puberty in which children begin to engage more with the local community (not at school/day care and not at home) via participating in social activities such as sports and recreation, youth clubs, and cultural events. As children move away from spending the majority of their free time at home or after-school care and begin to engage more independently with the society and the context surrounding them, doing more of this in the HL seems to provide critical opportunities for continued grammatical development even at relatively later ages.

The results from the **case-marking/passive comprehension** task revealed that Japanese HSs, in line with previous literature on different HS populations (Bayram et al., 2019; Chondrogianni & Schwartz, 2020; Hao et al., 2023), perform better with active than passive as well as canonical than non-canonical structures. Our findings, however, provide further insights into the development of these grammatical features by revealing that active canonical—a structure that is most frequent in the input—is acquired (with more than 80% accuracy) among even the youngest cohort in our dataset, while other structures that are non-canonical show chance-level performance in young childhood, yet gradually reach ceiling accuracy as HSs get older. This shows that while such structures are protracted in development, they do not necessarily remain at a state of perpetual arrested development. Such finding supports results of other studies (Daskalaki et al., 2022; Flores et al., 2017; Flores & Barbosa, 2014; Jia & Paradis, 2020) which report that language

competence/proficiency of child HSs can improve as they grow older, suggesting that HSs simply need more time than their monolingual counterparts to accumulate enough HL exposure to acquire some target structures. Our finding also provides further evidence to the idea that HSs may encounter more difficulties in structures that induce increased processing loads through reanalysis of theta-role mappings and rejection of dominant, canonical interpretation (Chondrogianni & Schwartz, 2020).

In a similar vein to the classifier comprehension results, Literacy and Proficiency modulated **case-marking/passive comprehension** accuracy regardless of age, Voice (active, passive), or Canonicity (canonical, non-canonical). That is, children who had richer literacy engagement and higher proficiency in the HL performed better on comprehension from young childhood through adolescence. Our results are in line with Bayram et al. (2019) who also found a modulatory effect of literacy/formal education on the production of passives in Turkish adolescent HSs, underlining a modulatory role of formal education in the standard variety for the development of some HL properties (Kupisch & Rothman, 2018; Rothman, 2007; Torregrossa et al., 2023).

Crucially, the interaction between Age and Home (Figure 13) underscores the fact that some experiential factors may matter more or less depending on the stages of development. In contrast to the patterns we see regarding the differential effect of HL engagement in community on the development of **passive production** accuracy, HL exposure and use at home modulates **case-marking/passive comprehension** accuracy up until HSs reach middle childhood (around ages nine to ten). This is probably because children tend to spend more time at home with their family members (especially the main caretaker) until they become increasingly autonomous enough to partake in activities outside of the home context. Thereafter, reductions in input and opportunities for meaningful engagement take place as a result of less time being spent at home and/or increased use of the majority

language (ML) at home as children become increasingly dominant in the ML and invite their ML relationships into the home. This may be a contributory reason why we see a modulatory effect of HL engagement in the community *after* puberty (for passive production), while exposure and use at home as well as the final education of the main caretaker better predicts case-marking/passive comprehension shortly *before* puberty.

Finally, as was the case for classifier comprehension and production, the **passive production** performance was predicted by Immersion and Proficiency. Although immersion experiences during holidays/vacations to Japan may at first seem like a trivial factor insofar as it accounts for a relatively small proportion of time in any given year, even when (self-rated) proficiency was controlled for, the significant effect of immersion experiences remained. In fact, immersion experiences may be a factor that has been often overlooked by previous studies, given that most existent questionnaires (such as Alberta Language and Development Questionnaire, Paradis et al., 2010 and The Language Experience and Proficiency Questionnaire, Marian et al., 2007) do not include any questions specific to language exposure and use during holidays (with the exception of Alberta Language Environment Questionnaire Heritage; Daskalaki et al., 2019). We interpret more HL exposure and use during holiday as a proxy for (richer) immersion experience not least because holiday travel to the homeland is not something all HSs have the same facility to do and because it provides a context in which more relationships with Japanese dominant speakers, especially peer-to-peer ones, can be forged/nurtured. The systematic influence of immersion experiences on HL grammar has also recently been shown in Chondrogianni and Daskalaki (2023), in which they found that visits to the country of origin (measured in weeks cumulatively over the past four years) predicted not only vocabulary skills, but also improved performance with interface (syntax-discourse) structures involving subject placement. Indeed, both short-term visits and long-term stays in the homeland have been shown to be beneficial

in reactivating grammatical structures in the L1 for adult attriters (Casado et al., 2023; Chamorro et al., 2016) and child returnees (i.e., children who return to the native language environment after spending significant time in a foreign majority language context; (Kubota et al., 2022; Treffers-Daller et al., 2016). Taken together, our findings highlight the determinism of receiving not only quantitatively more but also qualitatively variant HL exposure via naturalistic input for structures that may impose difficulties for HL children.

5. Conclusion

The present study investigated what underlying experiential factors predict the development of comprehension and production of classifiers, case-marking (in active and passive voice), and morphological passive structures in Japanese child HSs and whether various experiential factors modulate grammatical performance differentially from childhood through adolescence.

Our findings emphasize that the individual differences we see in HL grammar do not surface at random; rather, they are influenced by a complex interplay of experiential and environmental factors differentially across the lifespan. Notably, we found that certain factors, such as proficiency, immersion experiences, and engagement in literacy activities, systematically predict HL grammar acquisition (classifiers and case-marking/passives) in Japanese HSs. In the context of our HS sample in which the Socio-Economic Status of the families was relatively high (Mean = 3.07 from a scale of 0 to 4), factors such as immersion experience and literacy that goes beyond the norm of what HSs usually experience played a greater role in predicting HL grammatical development than factors such as HL input at home.

Most importantly, our work further contributes to the field by demonstrating that distinct factors affect HL development in different ways—some variables are more important than others in early childhood while other factors better predict HL development in adolescence.

On the other hand, some factors modulate HL grammar consistently across childhood through adolescence. On the practical side, these findings underscore the need for tailoring interventions, for example when constructing HL specific pedagogies, not only to individual needs and specific context, but also timing (age). Understanding that different variables hold varying significance at different developmental stages will allow relevant stakeholders such as parents, educators, and clinicians to adapt their approaches, ensuring continuous HL development from early childhood through adolescence.

References

- Atagi, N., & Sandhofer, C. M. (2015). Generic and specific numeral classifier input and its relation to children's classifier and number learning. *Psychology of Language and Communication, 19*(2), 101–127.
- Bayram, F., Rothman, J., Iverson, M., Kupisch, T., Miller, D., Puig-Mayenco, E., & Westergaard, M. (2019). Differences in use without deficiencies in competence: Passives in the Turkish and German of Turkish heritage speakers in Germany. *International Journal of Bilingual Education and Bilingualism, 22*(8), 919–939.
- Casado, A., Walther, J., Wolna, A., Szewczyk, J., Sorace, A., & Wodniecka, Z. (2023). Advantages of visiting your home country: How brief reimmersion in their native country impacts migrants' native language access. *Bilingualism: Language and Cognition, 1*–12.
- Chamorro, G., Sorace, A., & Sturt, P. (2016). What is the source of L1 attrition? The effect of recent L1 re-exposure on Spanish speakers under L1 attrition. *Bilingualism: Language and Cognition, 19*(3), 520–532.
- Chondrogianni, V. (2023). Individual differences differentially influence language domains and learning mechanisms. *Journal of Child Language, 50*(4), 823–826.
- Chondrogianni, V., & Daskalaki, E. (2023). Heritage language use in the country of residence matters for language maintenance, but short visits to the homeland can boost heritage language outcomes. *Frontiers in Language Science, 2*.
- Chondrogianni, V., & Schwartz, R. G. (2020). Case marking and word order in Greek heritage children. *Journal of Child Language, 47*(4), 766–795.
- Daskalaki, E., Chondrogianni, V., & Blom, E. (2022). Path and rate of development in child heritage speakers: Evidence from Greek subject/object form and placement. *International Journal of Bilingualism, 13670069221111648*.

- Daskalaki, E., Chondrogianni, V., Blom, E., Argyri, F., & Paradis, J. (2019). Input effects across domains: The case of Greek subjects in child heritage language. *Second Language Research*, 35(3), 421–445.
- Daskalaki, E., Elma, B., Chondrogianni, V., & Paradis, J. (2020). Effects of parental input quality in child heritage language acquisition. *Journal of Child Language*, 47(4), 709–736.
- De Cat, C., Kaščelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & Consortium, Q.-Be. (2023). How to quantify bilingual experience? Findings from a Delphi consensus survey. *Bilingualism: Language and Cognition*, 26(1), 112–124.
- Flores, C., & Barbosa, P. (2014). When reduced input leads to delayed acquisition: A study on the acquisition of clitic placement by Portuguese heritage speakers. *International Journal of Bilingualism*, 18(3), 304–325.
- Flores, C., Santos, A. L., Jesus, A., & Marques, R. (2017). Age and input effects in the acquisition of mood in Heritage Portuguese. *Journal of Child Language*, 44(4), 795–828.
- Hao, J., & Chondrogianni, V. (2021). Comprehension and production of non-canonical word orders in Mandarin-speaking child heritage speakers. *Linguistic Approaches to Bilingualism*.
- Hao, J., Chondrogianni, V., & Sturt, P. (2023). Heritage language development and processing: Non-canonical word orders in Mandarin–English child heritage speakers. *Bilingualism: Language and Cognition*, 1–16.
- Hao, J., Kubota, M., Bayram, F., Gonzalez A. J., Grüter, T., Li, M., & Rothman, J. (2024). Schooling and home language usage matter in heritage bilingual processing: Sortal classifiers in Mandarin. *Second Language Research*.

- Jia, R., & Paradis, J. (2015). The use of referring expressions in narratives by Mandarin heritage language children and the role of language environment factors in predicting individual differences. *Bilingualism: Language and Cognition*, 18(4), 737–752.
- Jia, R., & Paradis, J. (2020). The acquisition of relative clauses by Mandarin heritage language children. *Linguistic Approaches to Bilingualism*, 10(2), 153–183.
<https://doi.org/10.1075/lab.16015.jia>
- Kan, R. T. (2019). Production of Cantonese classifiers in young heritage speakers and majority language speakers. *International Journal of Bilingualism*, 23(6), 1531–1548.
- Kubota, M., Chondrogianni, V., Clark, A. S., & Rothman, J. (2022). Linguistic consequences of toing and froing: Factors that modulate narrative development in bilingual returnee children. *International Journal of Bilingual Education and Bilingualism*, 25(7), 2363–2381. <https://doi.org/10.1080/13670050.2021.1910621>
- Kupisch, T., & Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism*, 22(5), 564–582.
- Laleko, O., & Polinsky, M. (2013). Marking topic or marking case: A comparative investigation of heritage Japanese and heritage Korean. *Heritage Language Journal*.
- 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.
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4), 940–967.
[https://doi.org/10.1044/1092-4388\(2007/067\)](https://doi.org/10.1044/1092-4388(2007/067))

- Meir, N., & Janssen, B. (2021). Child heritage language development: An interplay between cross-linguistic influence and language-external factors. *Frontiers in Psychology, 12*, 651730.
- Mitrofanova, N., Rodina, Y., Urek, O., & Westergaard, M. (2018). Bilinguals' sensitivity to grammatical gender cues in Russian: The role of cumulative input, proficiency, and dominance. *Frontiers in Psychology, 9*, 1894.
- Mitrofanova, N., Urek, O., Rodina, Y., & Westergaard, M. (2022). Sensitivity to microvariation in bilingual acquisition: Morphophonological gender cues in Russian heritage language. *Applied Psycholinguistics, 43*(1), 41–79.
- Montrul, S. (2008). *Incomplete acquisition in bilingualism: Re-examining the age factor* (Vol. 39). John Benjamins Publishing.
- Naka, M. (1999). The acquisition of Japanese numerical classifiers by 2–4-year-old children: The role of caretakers' linguistic inputs. *Japanese Psychological Research, 41*(1), 70–78.
- Noji, J. (1985). *Yoojiki no Gengo Seikatsu no Jittai (The Language Development of a Child)* (1–1–4). Bunka Hyoron.
- Paradis, J. (2023). Sources of individual differences in the dual language development of heritage bilinguals. *Journal of Child Language, 50*(4), 793–817.
- Paradis, J., Emmerzael, K., & Sorenson Duncan, T. (2010). *Assessment of English Language Learners: Using Parent Report on First Language Development*. *Journal of Communication Disorders, 43*(6), 474–497.
- Paradis, J., Soto-Corominas, A., Daskalaki, E., Chen, X., & Gottardo, A. (2021). Morphosyntactic development in first generation Arabic—English Children: The effect of cognitive, age, and input factors over time and across languages. *Languages, 6*(1), 51.

- Rodina, Y., Kupisch, T., Meir, N., Mitrofanova, N., Urek, O., & Westergaard, M. (2020). Internal and external factors in heritage language acquisition: Evidence from heritage Russian in Israel, Germany, Norway, Latvia and the United Kingdom. *Frontiers in Education, 5*, 20.
- Rothman, J. (2007). Heritage speaker competence differences, language change, and input type: Inflected infinitives in Heritage Brazilian Portuguese. *International Journal of Bilingualism, 11*(4), 359–389.
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism, 13*(2), 155–163.
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Dunabeitia, J. A., Gharibi, K., Hao, J., Kolb, N., Kubota, M., & Kupisch, T. (2023). Monolingual comparative normativity in bilingualism research is out of “control”: Arguments and alternatives. *Applied Psycholinguistics, 44*(3), 316–329.
- Ruiting, J. (2016). *Language development in Mandarin heritage language children* [University of Alberta]. https://era.library.ualberta.ca/items/902a97f2-d04a-4c03-b50b-9876a7c76788/view/c012d8d8-cae2-49b9-8f99-2c9bb63c48c0/Jia_Ruiting_201608_PhD.pdf
- Sorace, A. (2011). Pinning down the concept of “interface” in bilingualism. *Linguistic Approaches to Bilingualism, 1*(1), 1–33.
- Soto-Corminas, A., Daskalaki, E., Paradis, J., Winters-Difani, M., & Al Janaideh, R. (2022). Sources of variation at the onset of bilingualism: The differential effect of input factors, AOA, and cognitive skills on HL Arabic and L2 English syntax. *Journal of Child Language, 49*(4), 741–773.

- Suzuki, T. (2005). Tan-itsu koobun no rikai kara saguru yooji no kakujoshi hattatsu (The development of Japanese case-markers observed through children's comprehension of single-argument sentences). *Gengo Kenkyu*, 132, 55–76.
- Torregrossa, J., Flores, C., & Rinke, E. (2023). What modulates the acquisition of difficult structures in a heritage language? A study on Portuguese in contact with French, German and Italian. *Bilingualism: Language and Cognition*, 26(1), 179–192.
- Treffers-Daller, J., Daller, M., Furman, R., & Rothman, J. (2016). Ultimate attainment in the use of collocations among heritage speakers of Turkish in Germany and Turkish–German returnees. *Bilingualism: Language and Cognition*, 19(3), 504–519.
- 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.
- van Osch, B., García González, E., Hulk, A., Sleeman, P., & Aalberse, S. (2019). The development of subject position in Dutch-dominant heritage speakers of Spanish: From age 9 to adulthood. *Languages*, 4(4), 88.
- 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.