Visual Heuristics for Verb Production: Testing a Deep-Learning Model With Experiments in Japanese

Chang, Franklin

Tatsumi, Tomoko

Hiranuma, Yuna

Bannard, Colin

(Citation)
Cognitive Science,47(8):e13324

(Issue Date)
2023-08

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/0100483317

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13324

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13324

Visual Heuristics for Verb Production: Testing a Deep-Learning Model With Experiments in Japanese

Franklin Chang,[a] Tomoko Tatsumi,[b,c] Yuna Hiranuma,[a]
Colin Bannard[d]

[a]*Department of English Studies, Kobe City University of Foreign Studies*
[b]*Graduate School of Intercultural Studies, Kobe University*
[c]*Language Development Department, Max Planck Institute for Psycholinguistics*
[d]*Department of Linguistics and English Language, University of Manchester*

Received 4 June 2022; received in revised form 16 June 2023; accepted 6 July 2023

Abstract

Tense/aspect morphology on verbs is often thought to depend on event features like telicity, but it is not known how speakers identify these features in visual scenes. To examine this question, we asked Japanese speakers to describe computer-generated animations of simple actions with variation in visual features related to telicity. Experiments with adults and children found that they could use goal information in the animations to select appropriate past and progressive verb forms. They also produced a large number of different verb forms. To explain these findings, a deep-learning model of verb production from visual input was created that could produce a human-like distribution of verb forms. It was able to use visual cues to select appropriate tense/aspect morphology. The model predicted that video duration would be related to verb complexity, and past tense production would increase when it received the endpoint as input. These predictions were confirmed in a third study with Japanese adults. This work suggests that verb production could be tightly linked to visual heuristics that support the understanding of events.

Keywords: Telicity; Tense; Aspect; Japanese; Morphology; Development; Deep learning

Correspondence should be sent to Franklin Chang, 9 Chome-1 Gakuen, Higashimachi, Nishi Ward, Kobe, Hyōgo Prefecture 651–2187, Japan. E-mail: chang.franklin@gmail.com

1. Introduction

A key part of acquiring a language is learning to encode tense/aspect distinctions in verb morphology. For example, in Japanese, the word *hashiru* ("run") can be made into the past tense form *hashitta* by changing the ending to *ta*, or the progressive form *hashitteiru* ("running") by changing the ending to *teiru*. In this paper, we seek to better understand how speakers learn to use the tense and aspect distinctions in Japanese, and specifically how they move from nonlinguistic visual cues to verb morphology. We report on experiments in adults and children, and use the resulting data to develop a computational model of Japanese verb production.

Tense and aspect are independent distinctions that can be fully crossed (simple nonpast *run*, progressive nonpast *is running*, simple past *ran*, progressive past *was running*). One prominent theory of how they are acquired across languages is the Aspect Hypothesis (Shirai & Andersen, 1995), which argues that there is a tradeoff between the choice of the simple past versus the progressive nonpast in language acquisition (in this paper, the word "past" refers to the simple past and the word "progressive" refers to the nonpast progressive). This past-progressive tradeoff depends on four different event types proposed by Vendler (1967). These are depicted in Fig. 1 and can be distinguished from one another as follows. Achievement and accomplishment events have an endpoint where the event can be said to be completed (telic), while activities and states do not have an endpoint (atelic). Achievement events center on change that occurs at the endpoint of the event (e.g., die). By contrast, accomplishments include the activity that leads up to the completion of the event (e.g., bake). Activity events involve some kind of motion or change without a fixed endpoint (e.g., run). States are persistent events without some physical activity (e.g., know).

The Aspect Hypothesis argues that telic events like achievements and accomplishments are more likely to appear initially in the past tense, while atelic activities are more likely to appear in the progressive. Shirai (1991; cited in Andersen & Shirai, 1994) found evidence that supported this link in the L1 English speech of three children. Further support was found in Japanese, where Shirai (1998) found that children used past *ta* forms early on with

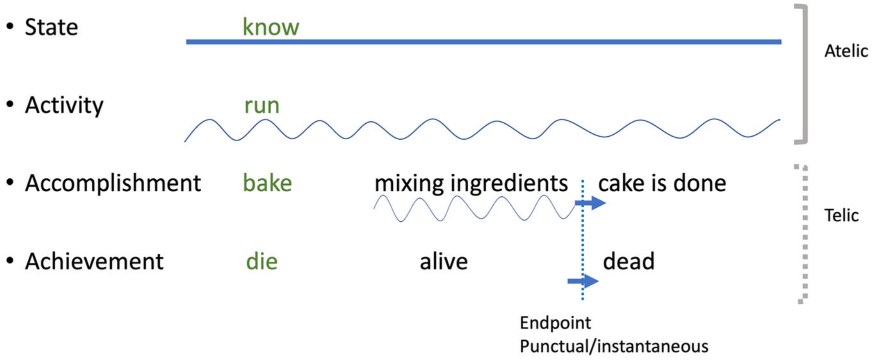


Fig. 1. Vendler's event types.

achievements and progressive *teiru* forms with activities, but at the same time, he did not find the predicted strong preference for using *ta* with telic accomplishments. In addition, Japanese speakers often used *teiru* with achievement verbs for resultative situations that are typically in the past (*hon-ga ochiteiru* means "the book has fallen," not the progressive "the book is falling now"; Shirai, 2000). Thus, the mapping between Vendler event types and tense/aspect is more complicated in Japanese (Shirai, 1998; Shirai & Kurono, 1998).

It is commonly assumed that verbs encode information that is useful for identifying event types and this is called the inherent lexical aspect of the verb. For example, the Japanese verb *suwaru* ("sit") has been classified as an achievement verb (Shirai, 1998), and when it is combined with *teiru* (*suwatteiru*) has the resultative meaning "I am seated." However, if one is viewing an old man who takes a few minutes to sit down, then it is possible to treat *suwatteiru* as an accomplishment verb with a progressive meaning ("he is slowly sitting down"). This suggests that the lexical aspect of a verb depends on a speaker's conceptualization of the event that is being described, but this is rejected in some of the early literature ("inherent lexical aspect are linguistic properties, and, … should not be confused with 'the properties of an actual situation.'"; Shirai & Andersen, 1995; van Hout, 2016). Previous work using corpus-based approaches where the situational context must be inferred also made it difficult to study how speakers conceptualized the events that they were describing. Here, we address this issue by manipulating visual information in the situation that is being described and see if these visual changes influence linguistic choices.

In this work, we propose that event types can be identified by using visual information about manner and goals (Levin & Hovav, 1991; Talmy, 1975, 1985). For example, the activity verb "running" focuses on the manner of motion of the agent (running involves the rhythmic back-and-forth motion of the agent's legs). On the other hand, the accomplishment phrase "running to the store" has an additional goal element, because the activity of running will end when the agent has achieved the goal of reaching the store. Achievements have a goal or result, but no manner component, while states have neither goal or manner. Manner can be identified visually from the motion of the agent's body parts (e.g., hands, feet), while the goal is visually related to how the agent interacts with external objects or entities (e.g., store). Furthermore, languages differ in how they map manner and goal/path/results into sentences. English often maps manner information to verbs and path information to nonverb elements like prepositions (e.g., "walking into a house"), while other languages such as Spanish or Greek often map path information to verbs and manner information to other elements like adverbs (Papafragou, Massey, & Gleitman, 2006; e.g., *entrar caminando a la casa* in Spanish would literally be "entering by-walking to the house"; Slobin, 1996). Japanese combines manner and goal/path information into complex verbs (e.g., *aruite haitta*, literally "walked and entered"; Matsumoto, 1996). Hence, it is important to consider visual cues for both manner and goals in order to understand event types in Japanese.

While there are many studies of how manner and goal/path/result cues influence the selection of verbs in different languages (Allen et al., 2007; Maguire et al., 2010), these studies have not focused on how these features influence tense/aspect morphology. Experimental studies of the acquisition of tense/aspect have used live demonstrations or videos to elicit verbs, but these studies have not controlled manner of motion and goal information

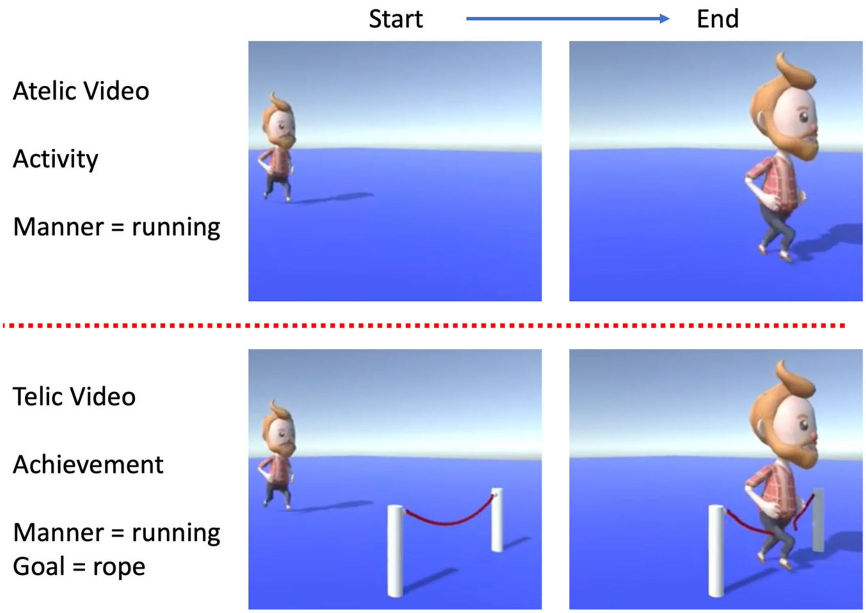


Fig. 2. Telic and atelic computer-animated scenes (Run).

(Kazanina & Phillips, 2007; McShane & Whittaker, 1988; Tatsumi, Ambridge, & Pine, 2018; Von Stutterheim, Andermann, Carroll, Flecken, & Schmiedtová, 2012). For example, Wagner, Swensen, and Naigles (2009) showed children a 6-second video of a woman picking flowers in the noncompleted condition, but in the completed condition video, picking-flowers manner information appeared for 2 seconds followed by an additional goal event where the woman showed a completed bouquet of flowers for 4 seconds. We think it is important to control manner and goal cues in studies of verb morphology, as studies have found that people spontaneously encode endpoints of visual events (Ji & Papafragou, 2022) and their processing of these events can be influenced by goal information (Mathis & Papafragou, 2022).

In the present work, telic and atelic computer-animated scenes were used to allow stronger control over the visual depictions of the actions. An example is shown in Fig. 2. In the atelic video, the man simply runs to the right side of the screen, while in the telic version, there is a finishing-line rope and the man breaks the rope. In both videos, the manner of motion of the man's legs is exactly the same. But in the telic scene, the rope provides a final goal for the running action. In addition, the rope breaking at the end of the event provides additional information that signals that an endpoint was reached. The first two studies examined whether Japanese adults and children will be more likely to use past tense verb forms when given goal/endpoint information. Those studies led to a computational model of Japanese verb production whose predictions were tested in a final study.

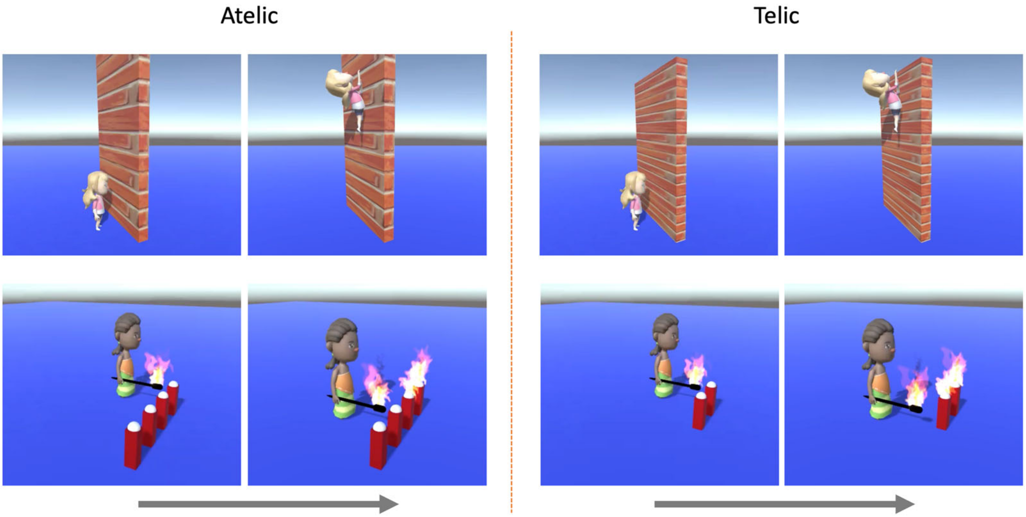


Fig. 3. Example videos of telic and atelic actions (Climb, Burn).

2. Experiment 1: Telicity in Japanese adults

Since very little is known about how visual information can influence the choice of verb morphology, the first study tested Japanese adults using 3D animations which had the same manner information, but which differed in the information related to the goal endpoint. Fig. 3 shows the start and end frames for videos depicting actions in either telic or atelic forms. The atelic version of each video showed an action performed by an agent character (e.g., climbing). The telic version of each video had the same manner of motion for the agent, but had differences in the other objects which created an endpoint for the event. For example in the videos on the top of Fig. 3, the climbing manner was the same, but the wall was shorter in the telic version, such that the agent reached their goal (the top of the wall) at the end of the action. Another example is the burn video at the bottom of Fig. 3, where the agent has lit all the torches in the telic video, but in the atelic video, there are still some extra unlit torches (possible future goals). Unlike the previous studies that used live-action videos or puppet enactments, the present study used computer-generated animations, where it was possible to present the same manner information in both videos (the movement of the hands and legs of the agent). If participants use goal information to understand telicity in the video, then they should use past tense more to describe the telic event than the atelic event. On the other hand, if tense/aspect morphology does not depend on visual cues like the length of a wall, then the same morphology will be used for both conditions.

2.1. *Participants*

The participants were 20 Japanese adults, opportunity sampled from the community of Kobe City University of Foreign Studies and other people resident around Kobe, Japan (some participants originally came from other regions in Japan). In a simulation-based power analysis assuming a medium effect size (9% difference between telic and atelic), a power level of 0.8 is achieved after 13 participants. The work was conducted with the formal approval of the Ethics Review Committee for Experimental Research with Human Subjects at Kobe City University of Foreign Studies.

2.2. Stimuli

Videos were generated within the Unity game engine (Unity Technologies, 2018) using characters from the Common People pack (Supercyan, 2018). Sixteen basic actions were created and situated within telic and atelic events (Climb, Jump, Sit, Enter, Paint, Kick, Throw, Switch Off, Switch On, Pull, Shake, Run, Split, Stack, Line Up, and Burn). To make it easier to understand the goal of the action, most of the videos had two attempts with the second one successful (e.g., the Throw action had two ball-throwing events and the second ball hit a target). Each action could appear with two characters from the eight total characters. Lists were created that first presented the 16 actions with one set of characters (alternating between telic and atelic) and then presented them again with the other character and the opposite telicity from the first version (this allowed telicity to be examined within participant). Four lists were created and participants were randomly assigned to each list. Two lists had one order of the actions and different telicity conditions for each action, while the other two lists were just a reverse order of the first two lists (counterbalancing the order of the actions). There were two practice items before the main stimulus items were presented. The videos stimuli, statistics, and modeling code are all available in an OSF archive (https://osf.io/9bhmn/).

2.3. Procedure

Participants participated in the experiment on a laptop using a program written in Processing (Reas & Fry, 2006). They were shown the videos and when each finished, a text box appeared and they wrote a description of what they had seen. To reduce explicit thinking about their verb choice, we asked people to also describe the color of the clothing of the characters in their responses.

The written responses produced by the participants were coded for tense and aspect morphology. In Japanese, tense and aspect are typically marked on the last verb in a sentence, so the final verb was extracted using an automated procedure that segmented based on particles, adverbs, and nouns. Then, past morpheme *-ta* (and its allomorph *-da*) and progressive morphology *-tei* (and its allomorph *-dei*) were coded. When the vowel *i* was omitted from *-teiru* forms (*janpu shiteru* "doing jumps"), they were marked as progressive nonpast (likewise for past progressive). Polite forms were also identified (e.g., *kirimashita* "switched off") and categorized for past and progressive. There were 414 past forms, 148 progressive forms, 3 progressive past forms, 58 simple nonpast forms, and 7 other forms (e.g., fragments, errors). No dialect forms (e.g., *-toru*) were produced. Since the Aspect Hypothesis focuses on the past-progressive tradeoff (Shirai & Andersen, 1995), simple nonpast or progressive past forms were excluded. The dependent measure for the analyses was a binary value for the production of simple past-tense forms as opposed to progressive nonpast forms (past = 1, progressive = 0). It should be remembered that although we manipulated the visual information in the

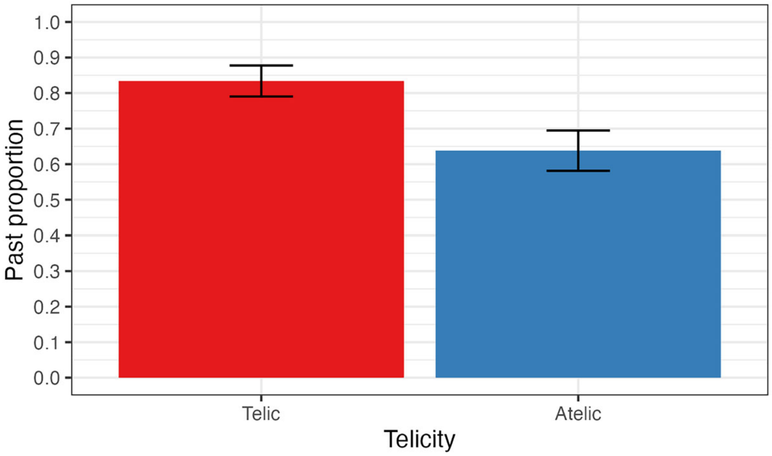


Fig. 4. Past form proportion of adult productions in Experiment 1 for telic and atelic videos.

videos in order to influence how participants conceptualized the event, we cannot infer from their verb morphology which particular meanings they were trying conveyed (e.g., *-teiru* can convey progressive, resultative, or iterative meanings).

2.4. *Results*

To examine whether the telicity in the videos influences verb morphology, we examined the likelihood of past forms being produced for telic and atelic videos (Fig. 4, all figures in this paper show 95% confidence intervals). A logistic mixed-effects model was fitted to past production with centered video telicity as a predictor. The maximal model that converged for the data had random effects on the intercept but no random slopes for either participants or video (Barr, Levy, Scheepers, & Tily, 2013). There was a significant effect of telicity, where videos that depicted endpoints were 20% more likely to be described in the past tense (excluding nonpast forms, $\beta = 2.12$, SE = 0.33, $\chi^2 = 52.35$, $p < .001$, fixed effects $R^2 = .076$; whole model $R^2 = .78$).

In developing this study, we expected that participants would use a small set of verb forms to describe each scene. For example, we expected that the running scene would primarily be described with the verb run, either in past *hashitta* or progressive form *hashitteiru*. In reality, however, participants produced a wide range of forms. For example, speakers used English loan words like *jogingu suru* (do jogging). There were also complex verbs such as *hashiri nukete kitta* (describing a telic event where the agent runs through the rope and cuts it into two), which is made up of several subverbs *hashiri* (run), *nukete* (through), and *kitta* (cut). The multiple subverbs complicate the use of event types for morphology, because “run” is an activity event type, “run through” is an accomplishment event type, and “cut” is an achievement event type.

To understand the range of verb forms that were used, we removed the past/progressive morphology from verbs (e.g., the running video was described by forms such as *hashi*,

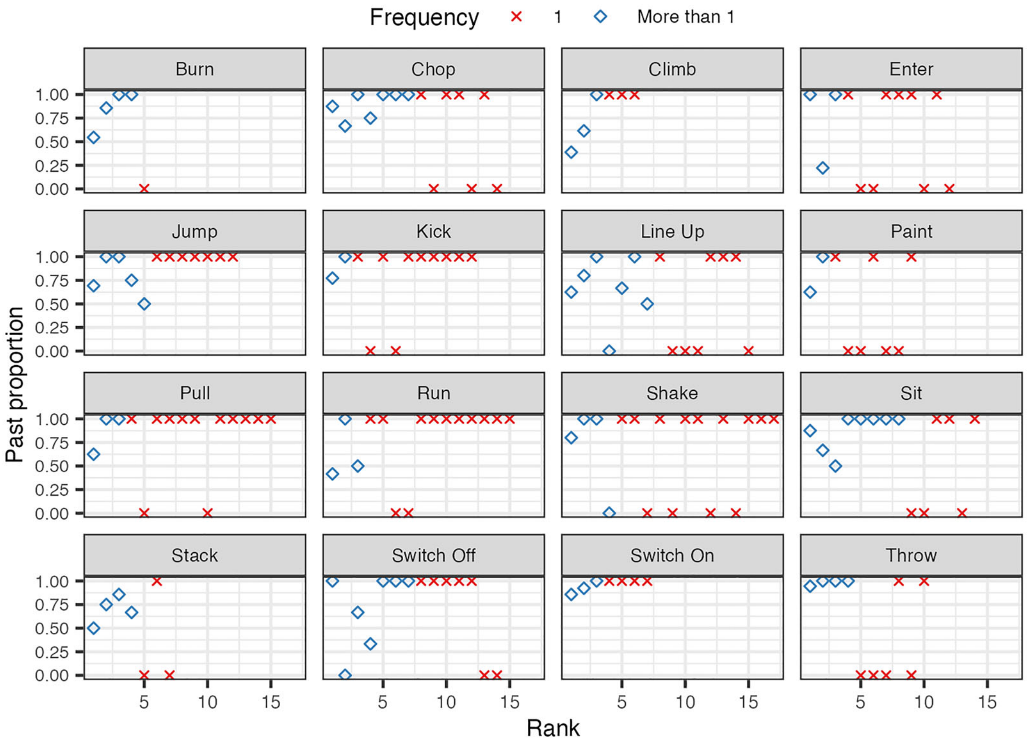


Fig. 5. Unique verb forms produced by adults in Experiment 1 for each video by past proportion and rank in our participant responses.

hashirinuke, *ranningushi*, or *kakenuke*) and found that there were 184 unique forms in this dataset (more than 11 forms on average for each action). For each of these forms, we computed its proportion in the past and its frequency in the participant responses, and the frequency within each action was used to assign a rank (most frequent = 1). Fig. 5 has a diamond or cross for each unique verb form with the rank and past proportion determining its position. As can be seen, participants produced multiple verb forms even with tense/aspect morphology removed. In addition, the figure shows that more than half of the verbs were used either with past endings only (111 verbs) or progressive endings only (40 verbs) compared to 33 verbs that occurred with both endings. The shape and color in Fig. 5 encode whether the form had a frequency of one (red cross) or a frequency that was greater than one (blue diamond). Evidence for an association between subverbs and tense/aspect comes from the many forms that occur more than once, but appear only in either past or progressive. Since acquisition theories would predict that speakers should choose the most frequent form that matches the meaning to be conveyed (Ambridge, Kidd, Rowland, & Theakston, 2015; Engelmann et al., 2019), the wide range of low-frequency subverb forms in the description of these simple actions is surprising and we will explore this issue in the later studies.

This study found that participants were more likely to use past tense when the video provided visual cues for goals (telic video). When the video did not have these cues (atelic video), speakers were more likely to use the progressive form. This occurred even though manner information that is often associated with the lemma in English (e.g., walk vs. run) was the same in both videos. In addition, we found that these videos elicited complex verbs made up of many subverbs (e.g., *hashiri nukete kitta*), which is due to the fact that Japanese speakers tend to combine manner (e.g., run) and goal/result (e.g., cut) information in verbs. We next explored whether young children can use the difference in goal information to select appropriate morphology.

3. Experiment 2: Telicity in Japanese children

The previous study provided evidence that past tense production in Japanese adults was sensitive to the visual properties of the videos. If this ability appears early in development, then it could play an important role in the acquisition of tense and aspect. There is some evidence that young children are quite sophisticated in their ability to understand action in visual events. When an entity moves toward an object, infants/toddlers will often treat the object as the goal of the movement (Gergely, Nádasdy, Csibra, & Bíró, 1995; Woodward, 1998). Furthermore, they appear to distinguish between events where the result is the intended goal of the action as opposed to events where it is an accidental outcome (Behne, Carpenter, Call, & Tomasello, 2005). Given these findings, children should be able to use visual cues for telicity to select morphological forms in a way that is similar to adults. But some studies have found that children are not able to use visual cues to learn information about verbs (e.g., verb-structure links, Twomey, Chang, & Ambridge, 2016), and if this is the case for verb morphology, then we might see no telicity difference in the children.

The present study was designed to test whether preschool children can recognize telicity in nonlinguistic events and can use appropriate morphology in production. This study used the same videos as the adults in Experiment 1. Shirai (1998) found evidence for the ability to use past and progressive morphology in event-type specific ways from 1 to 2.5 years of age. But since our videos involve actions and objects that may be difficult to understand, we tested older children between 3 and 5 years of age.

3.1. Participants

The participants were 41 children from a preschool in Kobe, Japan. This sample size was determined by the number of children who could be tested in a single preschool in the time available. Five children did not finish the study and were excluded from further analysis. Seven children produced nonpast or dialect forms, but produced no past or progressive forms. In the final dataset, there were 6 three-year-olds, 11 four-year-olds, and 12 five-year-olds. The work was conducted with the formal approval of the Ethics Review Committee for Experimental Research with Human Subjects at Kobe City University of Foreign Studies. The study was explained to parents and they were given a consent form to sign if they wanted their child to participate.

3.2. Stimuli

The videos for Experiment 1 were used in this study. The adults in Experiment 1 saw each video twice with different characters and different telicity conditions, but to reduce the task demands of the experiment, the children only saw each of the 16 actions once. Four lists were created and children were randomly assigned to each list. Two lists had one order of the actions and different telicity conditions for each action, while the other two lists were just a reverse order of the first two lists (counterbalancing the order of the actions). There were two practice items before the main stimulus items were presented.

3.3. Procedure

The procedure was designed to be a child-friendly way to elicit descriptions. First, a child was introduced to a penguin doll. Then, the penguin demonstrated two actions (sleeping and jumping), and the child was asked to describe the actions. Then, the child was told that they were going to watch some interesting videos and tell the penguin what they saw. The first two videos were practice trials, following which the experimental trials began. The child saw each video and was asked to describe it. If they did not respond, the penguin would say “Uh? What’s this? Could you tell me?” (*un, kore nani? oshietekurenai?*). If the child still did not respond, the penguin would say “Umm, I don’t understand it well … Could you tell me?” (*U-n, kore yoku wakannainaa, oshietekurenai?*). If the child did not respond to the second prompt, then the penguin would say “Umm, what could it be? Tell me” (*U-n, nandarou? oshiete*). If the child did not respond to the last prompt, then the penguin would suggest moving to the next trial.

The spoken responses of the children were coded using the same coding system as was used for the adults. Children tended to omit *i* from progressive forms (e.g., *aruiteru* instead of *aruiteiru*) which is a preferred form in informal conversation even for adults. Past and progressive endings were coded and other forms such as past progressive forms were excluded from the analysis (children produced 11 past progressive forms). Dialect forms were also excluded (past progressive *-totta, -dotta*; progressive nonpast *-ten, -den, -tou, -dou, -toru, -doru*). There were 82 simple past forms, 190 progressive nonpast forms, 11 progressive past forms, 97 simple nonpast forms, 37 dialect forms, and 75 other forms (e.g., fragments, errors).

3.4. Results

Our first question was whether the morphological choices that children made depended on the telicity cues in these videos. Fig. 6 shows the proportion of past forms for telic and atelic videos. A logistic mixed-effects model was fitted to past production with centered video telicity as a predictor. Random effects of participant on the intercept were included and the maximal model that converged for the data included no random slopes (fixed effects $R^2 = .022$; whole model $R^2 = .86$). There was a significant effect of telicity, where videos that

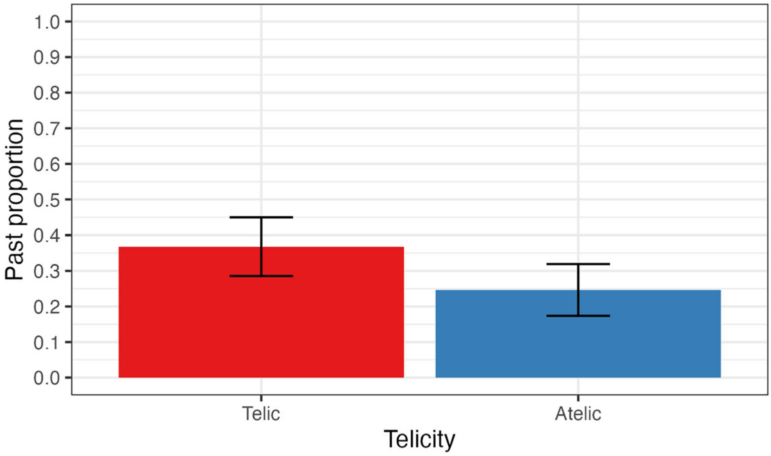


Fig. 6. Past form proportion of child productions in Experiment 2 for telic and atelic videos.

depicted endpoints were 12% more likely to be described in the past tense (excluding nonpast forms, $\beta = 1.41$, SE $= 0.46$, $\chi^2 = 8.56$, $p = .0034$). This demonstrates that 3- to 5-year-old Japanese children can distinguish telic and atelic actions based on visual cues and this telicity information is linked to past and progressive morphology such that they can select matching forms for each video.

As with the adult study, we examined the variation in the verb forms produced by the children. In addition to standard Japanese verbs (e.g., *keru*, kicking), children used English loanwords such as *kikku suru* "do kicking." They also used onomatopoetic verbs like *buranburan suru* (swinging back and forth) and complex verbs such as the verb *aruite tootte taoreta*, which is made up of the subverb *aruite* (walk), *tootte* (go through), and *taoreta* (fell). To examine this variation, we removed the past/progressive morphology from verbs and found that there were 106 unique forms in this dataset. For each of these forms, we computed its proportion in the past and its frequency in the participant responses, and the frequency within each action was used to assign a frequency rank (most frequent = 1). Fig. 7 has a cross or diamond for each unique verb form with the rank and past proportion determining its position. This figure shows that children produced a wide range of verb forms for these simple actions. In addition, more than half of the verbs used occur either in past only (31 verbs) or progressive only (54 verbs) compared to 21 verbs that occurred with both past and progressive endings. For example, children used *hipparu* (pull and stretch) only in the progressive form, while *taosu* (*knock over*) was only used in the past. So, even though these are young children, they appear to be sensitive to the association between verbs and past/progressive forms as has been found in corpora studies (Shirai & Andersen, 1995; Tatsumi, Chang, & Pine, 2021; Tatsumi & Pine, 2016).

To better determine whether children and adults differ in this task, a combined analysis of the adult and child data was performed. The adult and child counterbalance lists were exactly the same, except that the child lists were half as long to reduce the demands of the study. The

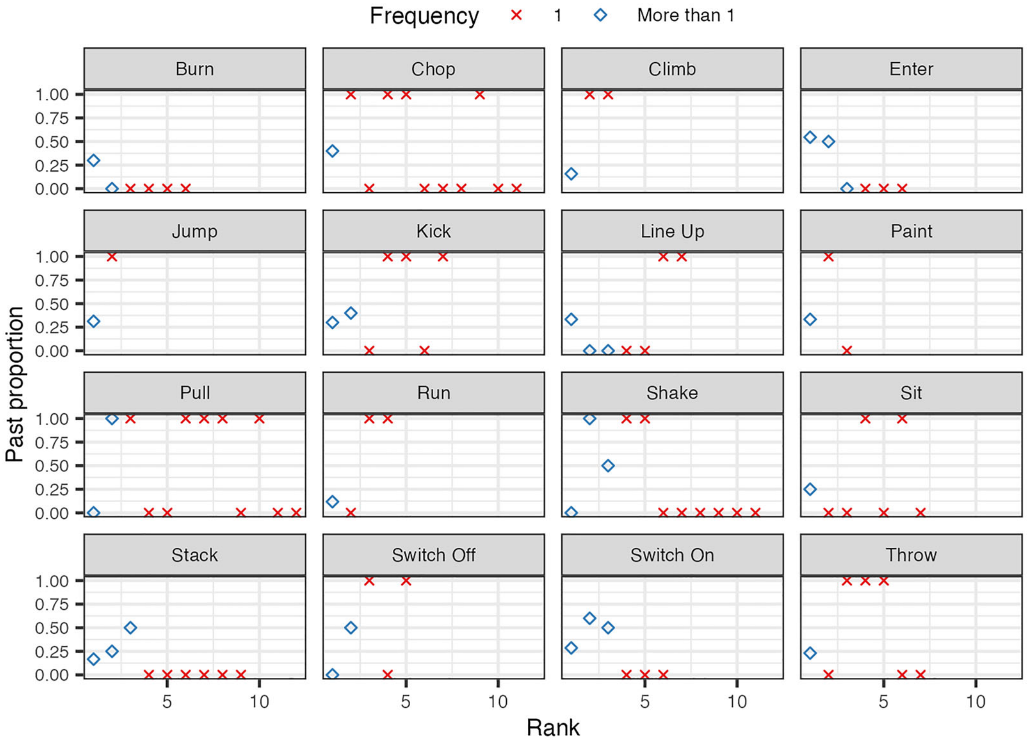


Fig. 7. Unique verb forms produced by children in Experiment 2 for each video by past proportion and rank in our participant responses.

child data and the first 16 actions for each adult were combined for this analysis. A logistic mixed-effects model was fitted to past production with centered video telicity crossed with centered age group (Fig. 8). Random effects of participant and video on the intercept were included and the maximal model had no random slopes (fixed effects $R^2 = .195$; whole model $R^2 = .85$). There was a main effect of age group ($\beta = 3.83$, SE $= 1.32$, $\chi^2 = 8.01$, $p = .0047$), because the adults preferred the past tense for these videos, while the children preferred the progressive. In this analysis, simple nonpast forms have been removed, but there was also a difference in those forms (adults use them 10% of the time, while children used them 45% of the time). There was also a significant effect of video telicity ($\beta = 1.51$, SE $= 0.34$, $\chi^2 = 24.02$, $p < .001$) and this did not interact with age ($\beta = 0.24$, SE $= 0.66$, $\chi^2 = 0.14$, $p = .713$; telic/atelic difference in past production was 15% in adults and 12% in children). The lack of an interaction with age indicated that the magnitude of the difference in past production between the telic and atelic videos did not vary strongly with development.

The first two studies have demonstrated that Japanese children and adults can use visual information about the intended goal to select tense/aspect morphology. In addition, we found that participants used a range of subverbs which complicates the identification of the event type of the verb. To better understand the complex processes necessary to parse visual events

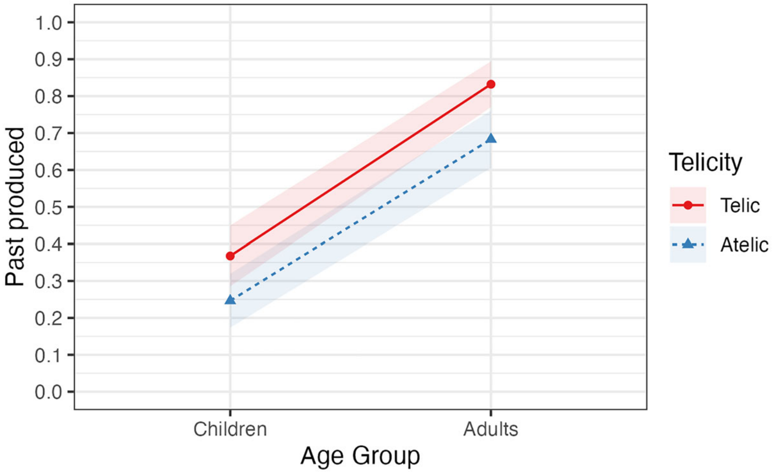


Fig. 8. Past form proportion of productions in Experiments 1 and 2 by telicity and age group.

into multiple components and map them to language elements, we developed a computational model that acquired the ability to produce Japanese verbs from visual input.

4. The Heuristics-Verb model of Japanese verb production

One of the unique features of the present study is that the videos were created in a 3D game engine, so it is possible to extract dynamic object movement information from these visual stimuli and use them to develop a computational model of Japanese verb production, which we call the Heuristics-Verb model. This model allows us to examine how a learner could learn to link Japanese tense/aspect morphemes to sequences of visual cues and how a range of subverbs can be generated during verb production. The model's predictions will be tested in a final study. There is a great deal of prior work modeling verb morphology, primarily in the context of the past tense debate. However, these models either did not include semantics (Plunkett & Juola, 1999) or included hand-crafted semantics (Joanisse & Seidenberg, 1999), and since little is known about the features that drive the production of Japanese subverbs, it is difficult to apply these approaches to this dataset. We, therefore, take a different approach.

In developing our model, we used deep-learning algorithms, which are used for state-of-the-art natural language processing systems (LeCun, Bengio, & Hinton, 2015). Deep learning makes use of the prediction error-based back-propagation algorithm that was popularized by Rumelhart, Hinton, and Williams (1986), which allows for the learning of complex mappings between inputs and outputs by using intermediate hidden layer representations. Deep-learning algorithms parallelize back-propagation to run on graphic processors, which has vastly increased their speed. This allows them to be applied to large datasets using many layers, which create deep networks (hence the name deep learning).

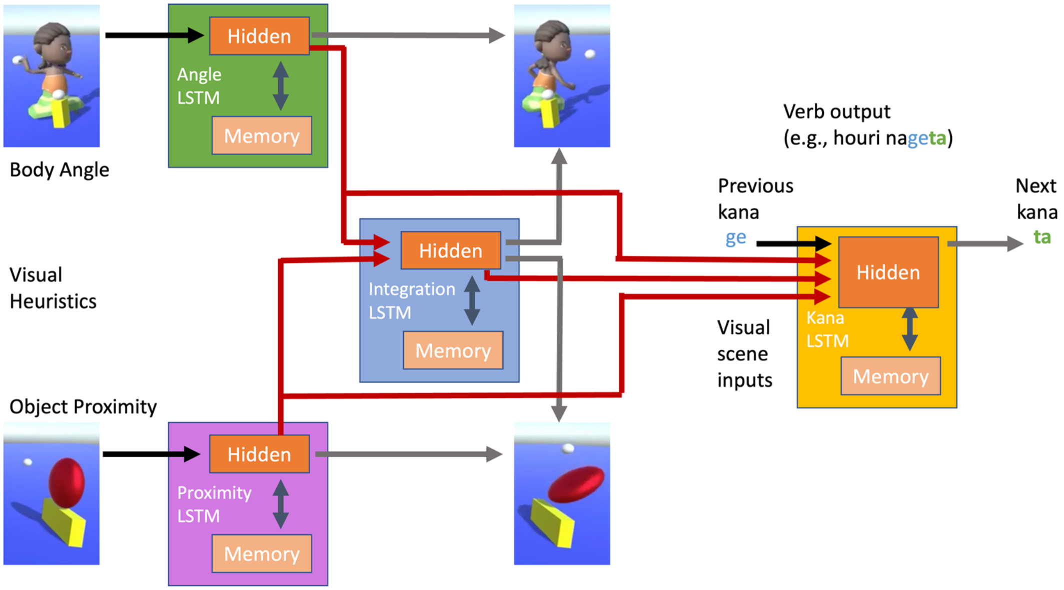
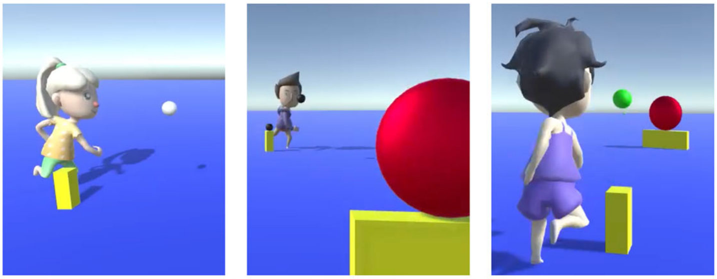


Fig. 9. Heuristics-Verb model architecture illustrated with a telic throwing action.

To produce a verb, the model will generate a sequence of kana, which are written symbols that correspond to syllable-like units in Japanese phonology. For example, one complex verb form was *houri nageta*, which combines the verb *houru* ("toss") with the verb *nageta* ("threw"). In the model, it would just be a sequence of kana *ho*, *u*, *ri*, *na*, *ge*, and *ta*. To generate these sequences, a recurrent network model will be used. An early type of recurrent model was the simple recurrent network (Elman, 1990), which mapped from the input at time *n* to the output at time *n+1* through a hidden layer with a context layer that stores a copy of the previous hidden layer representation (e.g., when given *ho*, the model would try to predict *u*). But simple recurrent networks have trouble learning long-distance dependencies and that has led to the development of more powerful recurrent algorithms such as Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997). Each unit in an LSTM model is made up of several components, one of which is a memory of the previous state. The unit can flexibly decide whether to switch its output depending on the new input or to continue to use the state in memory, and this allows it to easily learn long-distance dependencies. The Heuristic-Verb model incorporates a *Kana* LSTM submodel (right side of Fig. 9), which uses an LSTM to predict verb-specific kana sequences. The training inputs and targets for the Kana LSTM come from the verbs that the adults produced in Experiment 1. Each kana sequence began with a start symbol as input and the model generated a prediction about the next kana. The mismatch between the predicted kana and the actual next kana is the prediction error and this is used to adjust the weights in the model, so that in the future, the model will be better able to predict the next kana. This prediction-error-based learning mechanism derives support from studies where humans acquire transitional probabilities for syllables (Pelucchi,



Viewpoint	Side	Front	Back
Ball color	White	Black	Green

Fig. 10. Three throwing actions used in training action-understanding systems.

Hay, & Saffran, 2009) and generate prediction error signals which could be used for learning (Chang, Dell, & Bock, 2006; Emmendorfer, Correia, Jansma, Kotz, & Bonte, 2020; Fitz & Chang, 2019).

4.1. Visual heuristics for the Heuristics-Verb model

There are many deep-learning models for action understanding (Zhu et al., 2020). These algorithms typically map from pixel-based images in videos to an action classification. It is not clear what kinds of representations are learned in these complex opaque models, but their classification of actions can be sensitive to camera viewpoint and colors. For example, a model trained to classify the action of throwing with only the side and front views in Fig. 10 will have trouble classifying a view from the back. In contrast to these approaches, we propose that humans share similar view-independent primitives for understanding actions. This proposal is implemented through abstract visually based representations that support action understanding, which we call *visual heuristics* (Jessop & Chang, 2022). Here, we examine whether two types of heuristics (Body-Angle and Object-Proximity heuristics) can support the linguistic choices in Japanese verb production.

One set of heuristics was the Body-Angle heuristics. In the throwing event, the agent threw one ball and then another ball (Fig. 11). In the telic video, there was a red target that was knocked over on the second throw, while in the atelic video, there was no target. As the game engine was creating the video, we recorded the angle of the feet and hands with respect to the body of the agent. The bottom half of Fig. 11 shows the Body-Angle heuristics, where the right hand (Right Hand Angle heuristic) moved to pick up each ball and throw them (the other hand and feet moved as a result of the momentum of the throwing action). Thus, the Body-Angle heuristics captured manner of motion information that could help identify the lemma “throw.”

The cognitive plausibility of the Body-Angle heuristics is supported by work on biological motion using point-light displays. In these studies, lights are attached to parts of the body and videos are recorded in a dark room where only the movement of the lights can be seen (a point-light display for walking is shown in Fig. 12). Golinkoff et al. (2002) found that

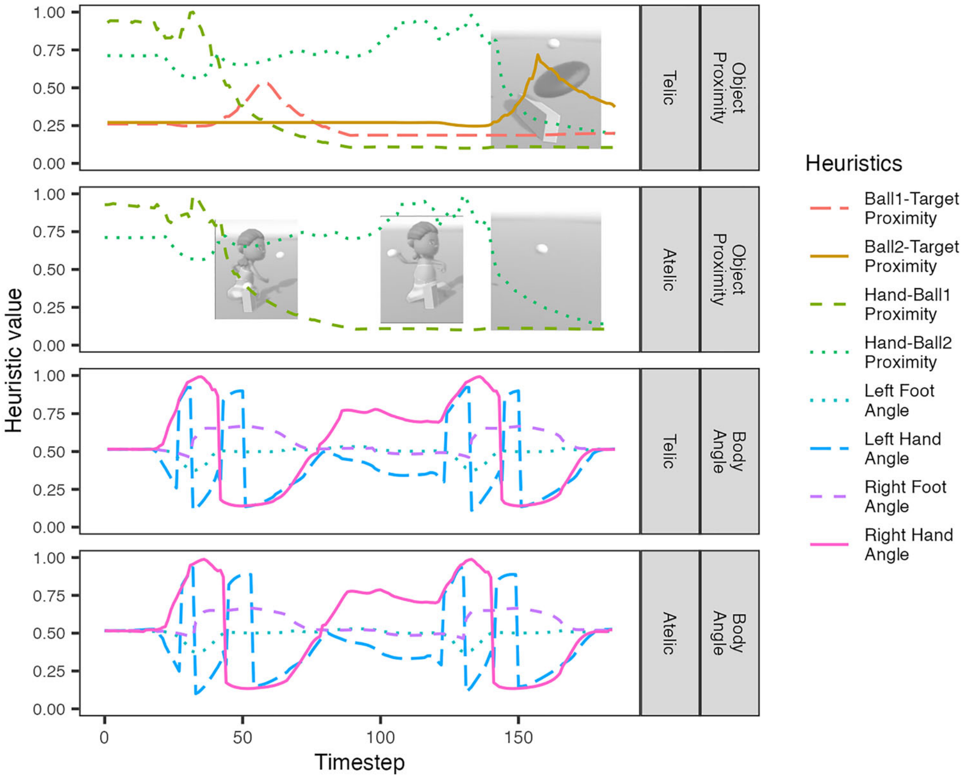


Fig. 11. Body-Angle and Object-Proximity heuristics for the throw action.

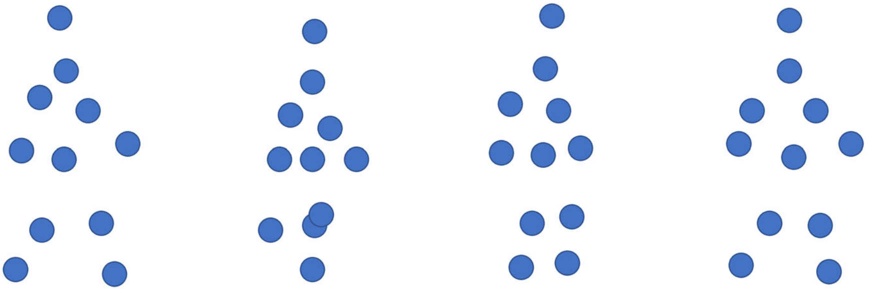


Fig. 12. Point-light displays: Four frames of a walking action.

3-year-olds can describe point-light displays with appropriate verbs (*walking*). Also using these videos, it has been found that 3-day-old newborns prefer biological motion over random motion (Bidet-Ildei, Kitromilides, Orliaguet, Pavlova, & Gentaz, 2014). Since point-light displays replace real-world objects (e.g., arms, legs) with points, the ability to recognize biological motion in a cloud of points indicates that there are specialized brain systems that can identify arm- or leg-like actions from just pure motion information. The Body-Angle heuristics are one way to implement these types of biological motion recognizers.

Another set of heuristics was the Object-Proximity heuristics (top half of Fig. 11). First, the Hand-Ball1 Proximity recorded that the right hand was in close proximity to the first ball as the agent held it. Then that heuristic fell as the ball is thrown. In the telic video, there was a heuristic recording of the movement of the first ball pass the red target (Ball1-Target Proximity rose around timestep 60). Then, the agent grabbed the second ball and threw it (Hand-Ball2 Proximity rose and then fell). The second ball hit the target and knocked it over (Ball2-Target Proximity rose around timestep 160). The atelic video did not have a target, and, therefore, there was no target-related proximity information. Therefore, the Object-Proximity heuristics allowed the model to recognize the target-related goal of the telic event.

The Object-Proximity heuristics assumes that humans encode the relative distance between objects. Evidence for this comes from work showing that humans and other animals encode distance to landmarks (Chan, Baumann, Bellgrove, & Mattingley, 2012; Waller, Loomis, Golledge, & Beall, 2000). These abilities appear early in development as children are better at finding objects near landmarks than they are without landmarks (Learmonth, Newcombe, & Huttenlocher, 2001). Furthermore, the ability of infants and toddlers to understand the movement of entities or objects toward a goal object requires a sensitivity to the change in proximity of the two objects (Gergely et al., 1995; Woodward, 1998). Thus, before language learning begins, humans are sensitive to goal-directed actions and the Object-Proximity heuristics are one way to implement this ability.

These examples above show that information about the manner of the action can be inferred from the Body-Angle heuristics and information about the goal can be inferred from the Object-Proximity heuristics. But given that there is a lot of information at every timestep in these actions, it is desirable to summarize this information. One way to do this is suggested by a recurrent network of action segmentation developed by Reynolds, Zacks, and Braver (2007), which encoded visual scenes by predicting the position of body parts at time $n+1$ from the position of body parts at time n. We used a similar approach in the present work by using an Angle LSTM to predict the Body-Angle heuristics, and another Proximity LSTM to predict the Object-Proximity heuristics (left side of Fig. 9). These LSTMs compress the useful cues for action prediction into a static representation in their hidden layers.

Since the Angle and Proximity LSTMs were isolated from each other, they could not learn regularities that involved both systems (e.g., the manner of motion of the hand predicted that the ball will make contact with the target). To allow the model to integrate information from the Body-Angle and Object-Proximity heuristics, a third Integration LSTM was also included (middle of Fig. 9). This LSTM took input from the hidden layers of the Angle and Proximity LSTMs (red arrow inputs) and developed a common representation in its hidden layer that was shaped by the need to predict both heuristics for the next timestep (gray arrow output). The Integration LSTM would be the system that encodes the event type of an action, because it would be able to learn the relationship between the extended manner action information in the Angle LSTM (e.g., the arm moving) and the goal/result information in the Proximity LSTM (e.g., the ball hitting the target).

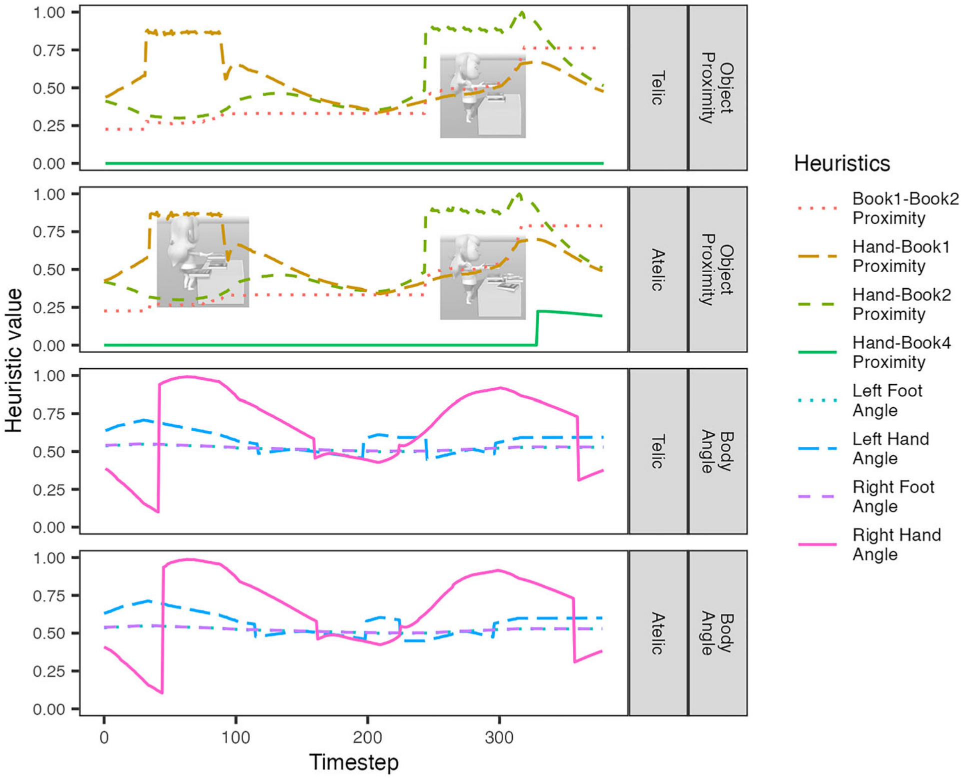


Fig. 13. Body-Angle and Object-Proximity heuristics for the stack action.

Once the video was fully processed, the hidden representation in the Angle, Proximity, and Integration LSTMs was used as visual input to the Kana LSTM. For example, for the verb *mukete nageta* ("throw toward"), it is likely that the Kana LSTM would depend on target-related goal information in the Proximity LSTM to produce *mukete* (toward the target) and the hand manner of motion information in the Angle LSTM to produce the subverb verb *nage* (throw). To know whether to use past tense *ta*, the model would likely use the combined representation in the Integration LSTM, which might encode that the throwing action was completed by hitting the target.

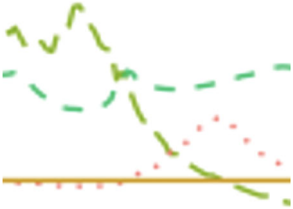
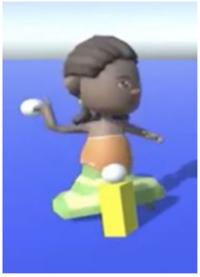
Most of the actions in the videos had clear endpoints that were signaled by relational object information. But there were four iterative events videos that did not signal endpoints in this way. One iterative event was the stacking video, where a girl stacked several books. Each time a book is put on the stack, one action is completed. But in iterative events, the progressive form "she is stacking books" refers to a sequence of these individual stacking actions. In these videos, the telic video showed that all possible goal objects had been acted on (all three books were stacked), while the atelic video showed that there were other objects that could still be acted on (only three of the six books were stacked). The Body-Angle heuristics for stacking are shown in Fig. 13, and it shows that the arm is moving back and forth to stack each book in the same way in the telic and atelic videos. But since the extra three books in the atelic version do not move or change at the endpoint, there is no clear cue that the endpoint of the sequence has been reached.

The difference in telicity in this video is due to the possibility of future stacking actions on the remaining three books in the atelic video. To simulate this within the present framework, we will make use of capacity limitations that exist in object tracking. Multiple-object-tracking studies have shown that people can track around four objects in parallel (Jessop & Chang, 2020; Pylyshyn & Storm, 1988). At the start of the event, we assume that object tracking focuses on the objects that are initially moved (e.g., Hand-Book1 Proximity, Hand-Book2 Proximity). Once these objects are stacked, only one tracker is needed to track this stack of books and hence object-tracking resources are available to track additional objects such as the three books remaining to be stacked (new goals). This is shown in the Object-Proximity heuristics for the atelic video in Fig. 13, where the fourth book is not tracked initially, but only after the second book has been stacked (the Hand-Book4 Proximity rises above 0 around timestep 330 to show that the system has started to track that book). A similar approach was used for the other iterative videos (Burn, Chop, and Line up). We will call this assumption the Adaptive Tracking assumption and we will test it in the final experiment.

The visual heuristics in the Heuristics-Verb model are unlike the binary features or categories that are used to categorize the whole event in linguistic approaches (e.g., telic/atelic, dynamic/static; Shirai & Anderson, 1995; Linguistic Features column in Table 1), because they provide dynamically changing visually based information about multiple objects in the scene. They are also more insensitive to low-level view-dependent properties of visual images such as camera position and color/shape that are used by deep-learning models of action understanding (Zhu et al. (2020); Visual Images column in Table 1). For example, some action understanding models may recognize a ball by its shape on one frame of video, but not know that a ball on the next frame is the same object. This means that they have some object category information (ball), but they do not always track object identity across frames (the same ball is moving across the screen). Object category information is useful for selecting verbs (e.g., balls tend to be thrown), but we have left it out of the present model, because it is desirable to have object-category-independent representations so that verbs can be used with novel arguments (e.g., *throwing an eggroll*). In addition, these two visual approaches assume that visual primitives such as low-level pixel inputs or higher-level object-based heuristics appear in infancy (e.g., Woodward, 1998), while linguistic approaches argue that features like telicity are discovered/constructed from language input later in development (van Hout, 2016).

The final difference in the approaches is in terms of the opaqueness of the theories. Linguistic approaches are transparent, because the features are created to directly reflect the relevant distinctions. Action understanding systems are a complex set of opaque components. For example, there are pose estimation systems that could generate manner of motion information like the Body-Angle heuristics (Zheng et al., 2022) and there are object-tracking systems (Wang, Chen, Yang, Hu, & Zhang, 2016) that could generate goal information like the Object-Proximity heuristics. But it is often difficult to determine what representations are being used in these complex black-box models. The present work provides a simpler model with just two heuristics that can be used to generate predictions about human behavior. Since it is not

Table 1
Three approaches for encoding event information for verb selection

	Linguistic features	Visual heuristics	Visual images
Lemma	Action concept (e.g., THROW)	Manner (Body angle) Goal (object proximity)	2d pixel array input
Subverbs **Tense aspect morphology**	? Event types (e.g., activity) Binary features (e.g., telic/atelic)		
Features	Transparent Not linked to vision Whole event Linguistically constructed	Semi-opaque View-independent Object identity Visual primitives	Opaque View-dependent Object category Visual primitives
Theories models	Aspect Hypothesis (Shirai & Anderson, 1995)	Heuristic-Verb model	Action understanding models (Zhu et al., 2020)

known if these visual heuristics are useful for language, we will first test whether they can be used to select Japanese verb lemmas, subverbs, and tense/aspect morphology in a way that is similar to the behavior in the first two studies.

4.2. *Model training and testing*

Training of the model involved the Body-Angle and Object-Proximity heuristics, which were collected as the animations were created at approximately 40 frames per second. The four Body-Angle heuristics represent the angle of the left hand, right hand, left foot, and right foot with respect to the body in the sagittal plane. For each action, the maximum and minimum values for both telic and atelic videos were computed and these were used to normalize the values between 0.1 and 1. Loss of object tracking was signaled by setting the feature to zero.

The Object-Proximity heuristics differed depending on the objects in each video. First, the raw distance in the game engine was computed at each timestep for the right hand and each object in the scene and likewise for the right foot and these objects. In addition, the raw distance for all pairs of external objects in the scene was also collected. Then for each of these distance features, the maximum and minimum values were computed separately for the telic and atelic videos and the difference between them (the span) was computed. The six heuristics with the largest span values were kept and these heuristics were used for both telic and atelic events. Reducing to six features in this way allowed us to focus on the heuristics that were the most distinctive, but also reduced variability between the actions in terms of the number of heuristics. Consistent with psychophysical work on distance perception (Posner, Goldsmith, & Welton, 1967; Stevens, 1957), the proximity heuristics were computed by taking the negative logarithm of the distance and normalizing the values between the maximum and minimum values for each action. This value was squeezed between 0.1 and 1, so that zero could be used to signal a loss of object tracking. This approach to computing proximity meant that there was a greater sensitivity to small distances (a small distance is important for distinguishing between “on the table” vs. “above the table”).

Deep-learning models are trained in parallel on multiple graphic processing units and typically have input sequences of the same length. Therefore, the action input sequences were made to be the same length as the longest sequence of 382 timesteps by adding zeroed Body-Angle and Object-Proximity heuristics to the earlier timesteps. The target action sequence was created by shifting the input action sequence by one timestep. The Angle LSTM in Fig. 9 took the Body-Angle heuristics at one timestep (black arrows) and predicted the Body-Angle heuristics at the next timestep (gray arrows), and the Proximity LSTM did the same for the Object-Proximity heuristics. The Integration LSTM took the hidden representations of the Angle LSTM and the Proximity LSTM at one timestep and predicted the target Body-Angle and Object-Proximity heuristics for the next timestep. Each model used the error in its heuristics prediction to adjust its weights.

The language inputs for the Kana LSTM were created from the verbs produced by the adults in Experiment 1. The written forms in that study were converted into kana and then presented to the model using the *kakasi* function in the Nippon R library (Tanimura, Takahashi, Baba, & Nokubi, 2018) as well as some manual conversion. The first input was a start symbol that signaled that production should start and this was followed by a sequence of kana for a particular verb (*START*, *ho*, *u*, *ri*, *na*, *ge*, *ta*). After the last kana, a period was added to signal the end of the kana sequence. The kana target was a shifted version (*ho, u, ri, na, ge, ta,.*) so that the model was predicting the next kana or period. To make the sequences the same length, a separate empty symbol was added to make the sequences the same length as the longest verb in the dataset (17 kana).

The action sequences and the language inputs were combined together on the assumption that people often learn or use verbs for describing actions that they have just seen. The 382 timesteps of the action sequences were followed by the 17 timesteps of the kana language sequence to create sequences 399 timesteps long. During the action sequence part of the input, the kana inputs and targets were set to zero, and likewise, during the kana sequence, the heuristics in the action sequence were set to zero. To give the model some experience with a delay between the action sequence and kana production, there were random numbers of timesteps (from 1 to 3) between the end of the action sequence and the start of the language sequence which were set to the empty symbol for both kana sequences and zeros for the action sequences. The input was shuffled in groups of 512 patterns and then batched into sets of 50 for training. This created variation in the order during training to avoid overfitting a particular batch.

The Angle, Proximity, and Integration LSTMs had 10 hidden units each and used a mean-squared error loss. The Kana LSTM had 40 hidden units and used a categorical cross-entropy loss, which created a winner-take-all bias for kana selection. The hidden representations in the LSTMs were reset at the start of each action/verb sequence. Units were initialized with glorot uniform initialization (Glorot & Bengio, 2010). An adaptive learning rate optimization algorithm with both momentum and scaling was used (Adam; Kingma & Ba, 2017). Ten models with random initial weights were trained for 2000 epochs and tested every 50 epochs. Simulations were run using Tensorflow 2.4.1 on an NVIDIA Titan X Pascal GPU and took about 1.5 h to train the 33,412 parameters in each model.

Although training involved mapping between particular action inputs and kana targets, testing involved a generative procedure where there was no target form. First, the action feature inputs from the first 382 time steps were presented to the model to simulate the viewing of the video by the experiment participants. Then, the start symbol was presented as input to the kana production LSTM. The activation on the output kana layer represented a probability distribution for the next kana, so a single kana was sampled from this categorical distribution and this was treated as the kana produced by the model. Before sampling, the distribution was divided by a temperature value of 0.5, which biased the model away from lower probability kana. The sampled kana was then passed into the model as the input for the next time step and the model generated its prediction for the next kana output. For example, after seeing the throw video, the probability of *na* is higher than the probability of *ho* as the first kana, because of the distribution of verb forms in the language (many throw verbs start with *na*). This generative procedure was applied to all of the inputs in the training set and this yielded a kana string for each action. The produced verb was extracted by removing all characters after the period was produced. These produced verbs were coded for whether they had past or progressive endings. After removing the past or progressive morphology, we coded whether the remaining string matched at least one form produced by the adults in Experiment 1 for each action and this was called the adult-like coding. For example, if the model produced *houri nageteiru*, this would be classified as adult-like due to the overlap with *houri nageta*, even though the *teiru* form never occurred with this verb in training.

4.3. Producing multiple different verb forms for each action

One of the goals of this modeling work is to explain how Japanese speakers produce a range of different verb forms for each action. In the generative procedure used in testing, kana were selected probabilistically from the kana output distribution and these kana were sequenced together to create different verb forms. This approach means that the model was generating different forms based on kana probabilities given the action in the video, but there was no meaning cue that signaled which particular verb forms to use. But presumably, speakers select particular complex verbs, because they want to highlight different components of the action. For example, the six-kana verb *mukete nageta* encodes the early throwing part of the event *nageta* (threw) as well as the late part of the event *mukete* (toward the target). But if they

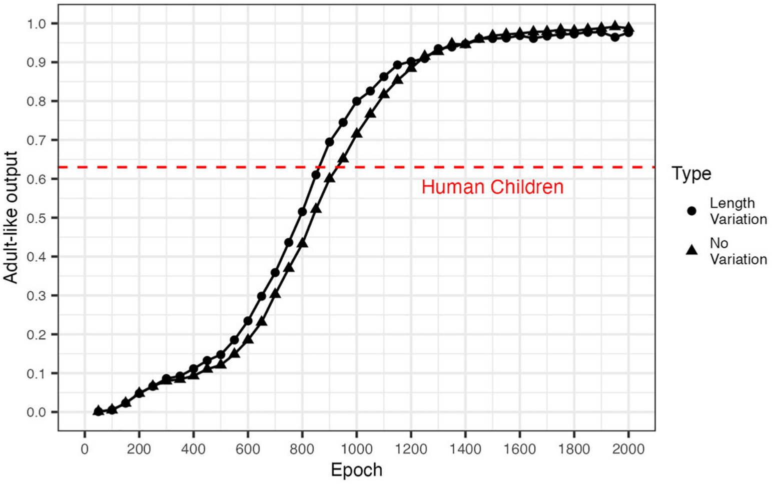


Fig. 14. Adult-like proportion by training epoch in Length Variation and No Variation models.

just focused on the hitting of the target, they could use a shorter three-kana verb like *ateta* (hit the target). Thus, we hypothesized that speakers choose different verb forms in order to convey different components of the meaning of the action. There is some support for this in that experimental studies on iconicity have found that novel names evolve to match the length of the corresponding action (Jones et al., 2014).

To simulate this effect of action meaning in the model, we developed a *Length Variation* version of the model's input which varied the length of the action input in response to the length of the produced verb in kana. For example, the full length of the action input for the Enter video was 184 timesteps and this was paired with the longest verb for this video, which was the 16 kana verb *aruite haitte ikouto shiteiru*, which includes subverbs *aruite* (walk), *haitte* (enter), and *ikou to shite* (try to go). The shortest verb for this video (3 kana *hairu* "enter") would be paired with half of the full-length action input (92 timesteps). Verbs that were intermediate between these two extremes would have action sequence inputs that were interpolated between them (the 7 kana verb *hairouto shita* "tried to enter" would have an action sequence of 120 timesteps). This model can use the length of the action input to predict the number of subverbs to produce and this instantiates the idea that verb complexity is related to the amount of meaning that speakers want to encode. The first set of analyses compared this *Length Variation* model with the original version *No Variation* model (each verb in an action had the same action sequence).

The first analysis examined how often the model produced verbs that adults had produced in Experiment 1. Fig. 14 shows the proportion of adult-like production every 50 epochs in training for the Length Variation model and the No Variation model. Using the data from the fully trained model (epoch 2000), a logistic mixed-effects model was fitted to adult-like production with centered length variation as a predictor. Random effects of model participant and video on the intercept were included and the maximal model that converged for the data had length variation random slopes for both model participant and video (fixed effects R^2 = .038; whole model R^2 = .16). The No Variation model is slightly higher at the end of training $\beta = -0.77$, SE = 0.2, $\chi^2 = 10.56$, $p = .0012$, but both models were able to produce adult-like verbs more than 97% of the time. To model developmental behavior, it is necessary to identify the model that best approximates human children. The children in Experiment 2 produced adult-like productions 63% of the time (red dashed line in Fig. 14), and that level is reached at around 900 epochs in the Length Variation model and 950 epochs in the No Variation model.

The easiest way for the model to acquire this high level of adult-like output is to learn only the most frequent verb form for each action. The model can focus on learning this single form and at test, there is no competition between different forms. But this would not match the variation in verb forms seen in the Japanese speakers in the two experimental studies. To see if the model has acquired a human-like distribution of verbs, we compared the model's input distribution with the output distribution of the Length Variation or No Variation models. The frequency of each unique verb form in the model's input was computed and the rank was assigned (most frequent = 1). Those rank numbers from the model's input were attached to the verb forms in the model's output (only adult-like verb forms that were in the input were considered in this analysis).

Fig. 15 shows the average log frequency by the rank of each verb form within each action for the model's original input, Length Variation model output, and No Variation model output. The figure shows that both the Length Variation model (red crosses) and the No variation model (blue triangles) are producing multiple verbs for each action and the log frequency of these forms is close to the frequency in the Original Input (green circles). Visually, the main difference is that the No Variation model overproduces some of the low-frequency verbs for some actions (e.g., Enter, Sit).

To determine which model best fits the human distribution, we computed the absolute value of the difference in frequency between the model and the input frequency for each verb form (if the model did not produce the form, then its frequency was 0). Using the data from the fully trained model (epoch 2000), a mixed-effects model was fitted to this absolute-value input-output frequency difference for each verb in each action with centered variation (length or no variation) as a predictor. Random effects of video on the intercept were included and the maximal model that converged for the data had no random slopes (fixed effects R^2 = .005; whole model R^2 = .1). The Length Variation model produced verb distributions that were closer to the input distribution than the No Variation model $\beta = -0.4$, SE = 0.12, $\chi^2 = 12.1$, $p < .001$). So, while the No Variation model is slightly better than the Length Variation model in the production of adult-like forms, it is producing some forms more frequently than the human input. The Length Variation model provides a better match to human verb distribution and this model will, therefore, be used for the remaining tests.

4.4. Linking endpoints and morphology

The model learned to encode the action sequences and Japanese verb phonology. Here, we examine whether the model can exhibit the telicity difference in verb morphology found in

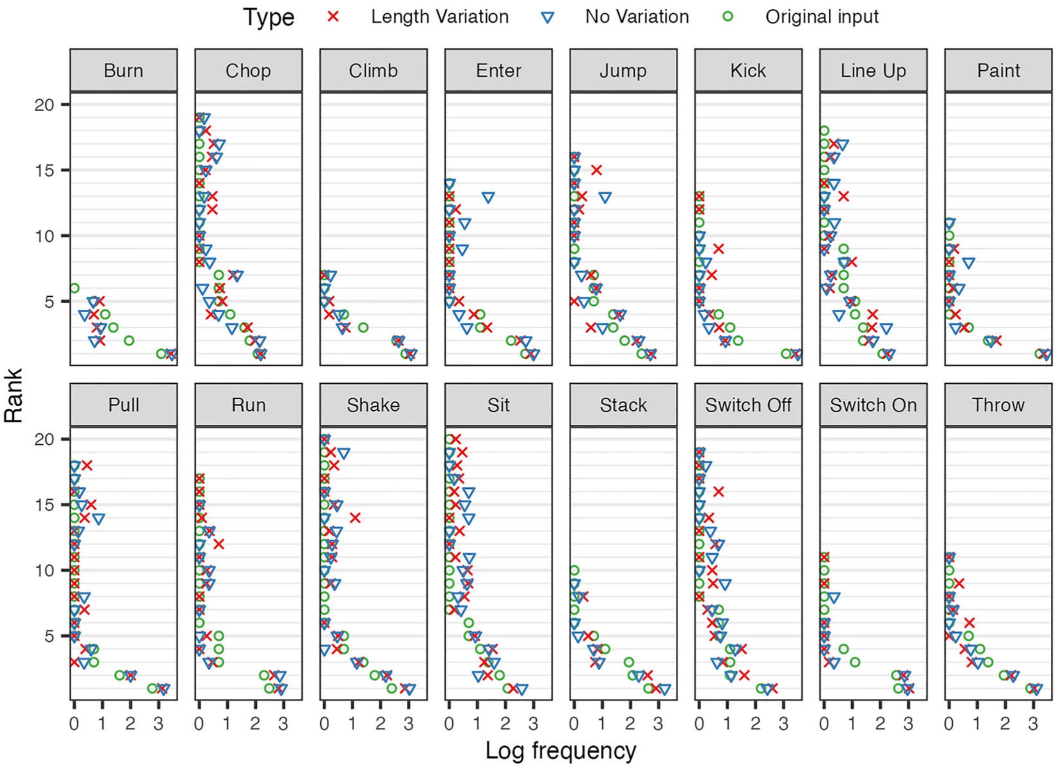


Fig. 15. Rank-log frequency distribution of verb forms in model's input, Length Variation output, and No Variation output.

the adult and child study. One possibility is that it would take time for the model to acquire an adult-like understanding of endpoints and this predicts that the child model might not show a telicity difference. But in the human experiments, the adults and children showed a similar telicity difference, which suggests that the ability to visually understand endpoints occurs early in development. To test these predictions, we used the Length Variation model at the end of training (epoch 2000) as an adult model and the model at epoch 900 as the child model based on the similarity to the children in the production of adult-like forms that was shown earlier. The input had both telic and atelic action sequences and the model's output was coded for past or progressive morphology.

Fig. 16 shows the proportion of past forms for the model and human data. The model prefers the past form early on in development and this is because the input is from the adult data in Experiment 1, where past production was dominant. Over development, the distinction between telic and atelic events appears to grow. To examine this, a logistic mixed-effects model was fitted to the past proportion for the child and adult models with centered telicity and age crossed. Random effects of model participant and video on the intercept were included and the maximal model that converged for the data had telicity and age main effect random

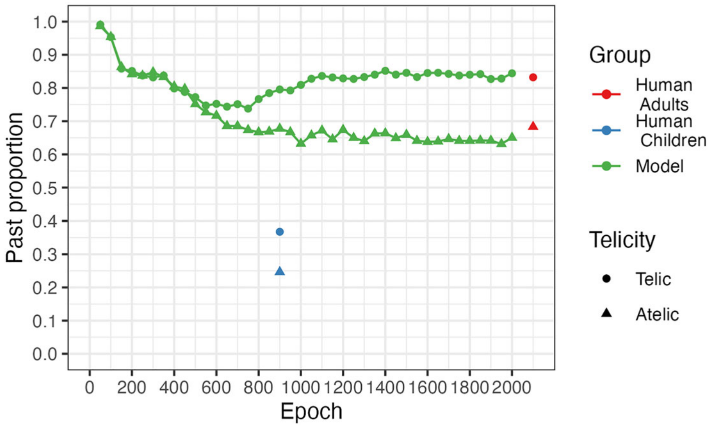


Fig. 16. Human and model results by telicity over development.

slopes for video and telicity random slopes for model participant (fixed effects $R^2 = .067$; whole model $R^2 = .25$). Telic inputs yielded more past descriptions than Atelic inputs $\beta = 1.03$, SE $= 0.33$, $\chi^2 = 7.12$, $p = .0076$. There was no main effect of age, but there was an interaction of age and telicity $\beta = 0.61$, SE $= 0.1$, $\chi^2 = 39.87$, $p<.001$, which is due to the effect being bigger in the adult model than the child model. However, when we test each model separately, there is still an effect of telicity at both epochs tested (child model, $\beta = 0.71$, SE $= 0.34$, $\chi^2 = 3.92$, $p = .048$; adult model, $\beta = 1.48$, SE $= 0.43$, $\chi^2 = 9.03$, $p = .0027$). Thus, the model evinced an early telicity difference and this is likely due to the fact the model does not need to abstract telicity from pixel-based visual input, but instead can quickly encode it from the visual heuristics.

Although the model exhibited a difference in telicity, it is not clear how it actually recognized endpoints in the action sequence. This is because the model was tested using the same stimuli used in the earlier human studies, where telicity was manipulated by changing external objects in the scene (e.g., finish rope in the running event). Since these objects were present throughout the event, the model might encode telic and atelic videos differently before seeing the endpoint. In general, it is challenging to find the endpoint, as it varies for different events (e.g., reaching the top of the wall) and there is no invariant cue for identifying them. To test if the model was sensitive to the endpoint, we tested the model with videos that stop before the endpoint. First, the timesteps in the action sequence were examined and the endpoint in each telic event was identified (e.g., the target being hit by the ball). Then, a new testset was created by taking the original telic and atelic training stimuli and replacing the Object-Proximity heuristic input from one timestep before the endpoint until the end of the action sequence with zeros. This Before Endpoint testset provided proximity heuristics about external objects that were potential goals for the event in the telic videos (e.g., target), but the changes at the endpoint were absent (e.g., target falling over). This testset was compared with the After Endpoint testset, which was just the original training stimuli that included the endpoint.

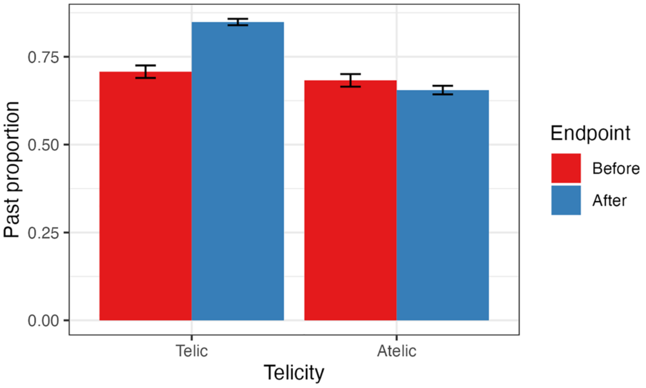


Fig. 17. Past form proportion of model productions depending on telicity and endpoint visibility.

Fig. 17 shows the proportion of productions that had past form for telic and atelic action inputs when the action ended before or after the endpoint. A logistic mixed-effects model was fitted to the past proportion with centered telicity and endpoint crossed. Random effects of model participant and video on the intercept were included and the maximal model that converged for the data had telicity and endpoint main effect random slopes for both model participant and video (fixed effects R^2 = .071; whole model R^2 = .31). Telic inputs yielded more past descriptions than atelic inputs $\beta = 0.85$, SE = 0.32, $\chi^2 = 7.46$, $p = .0063$ and the after endpoint condition yielded more past descriptions than the before endpoint condition $\beta = 0.31$, SE = 0.27, $\chi^2 = 5.38$, $p = .020$. However, these effects need to be considered in terms of the interaction of endpoints and telicity $\beta = 0.96$, SE = 0.09, $\chi^2 = 119.57$, $p<.001$. The interaction is due to an effect of telicity in the after endpoint condition $\beta = 1.54$, SE = 0.5, $\chi^2 = 7.64$, $p = .0057$, but not in the before endpoint condition $\beta = 0.23$, SE = 0.16, $\chi^2 = 2.13$, $p = .144$. This demonstrated that motion toward the goal object (e.g., finishing line) by itself was not the main cue that the model used to select past tense—instead the model was waiting for the changes at the endpoint as the main tense cue. This endpoint sensitivity appears to depend on the Integration LSTM, because it allowed the model to learn that the changes at the endpoint in the Object-Proximity heuristics (e.g., hitting the target) are the result of the action that began earlier in the Body-Angle heuristics (e.g., moving the arm).

One important difference between the After and Before Endpoint conditions was that the After Endpoint conditions were trained, while the Before Endpoint conditions were novel inputs, because the model was always trained with the proximity heuristics after the endpoint. Therefore, the production of adult-like output in the After Endpoint condition was 98%, while it was only 33% in the Before Endpoint condition. This reflects the fact that while deep-learning models are quite good at learning complex mappings, they are also quite sensitive to small changes in the input. This is a well-known problem for models trained with back-propagation (Chang, 2002; Marcus, 1998). The low accuracy here could be addressed by training with inputs with omitted data that is similar to the Before Endpoint testset. But since our goal here was to understand whether the model could learn to focus on the endpoint without any special training, we have not manipulated the training to increase adult-like production.

This Heuristics-Verb model tries to explain how dynamic movement in visual scenes can be used to select verb lemmas, subverbs, and verb morphology. It uses the LSTM learning algorithm to correctly produce quite long verbs such as *aruite haitte ikouto shiteiru* ("trying to go walk and enter"). Rather than just sampling kana randomly to make verbs, a better fit to human verb distributions was found when the model could use information about the amount of meaning to convey (Length Variation Assumption). The model also used past tense more for telic events and progressive more for atelic events as in the human data. This telicity distinction appeared early in development, because the model started with abstract view-independent heuristics and became sensitive to changes at the endpoint in telic events. The success of this model provides the first demonstration that, without any knowledge of the categories of the objects in the scene (e.g., book, ball), just two visual heuristics were sufficient to capture various aspects of Japanese verb production.

5. Experiment 3: Testing the predictions of the model

The first two experiments demonstrated that children and adults are sensitive to visual cues for telicity and a large variety of subverbs are used in Japanese. The modeling work showed that the ability to distinguish telic and atelic events could be explained by visual heuristics related to the endpoint of the event. To see if Japanese speakers are similar to the model in their sensitive to the endpoint, we took the telic videos and created two versions—an *Endpoint Seen* version where the video continued after the endpoint and an *Endpoint Unseen* version, where the action stopped before the endpoint. In this case, both videos have information about possible goals (e.g., the finish line in the running video), but the critical change at the endpoint (e.g., breaking of the rope) is not available in the Unseen video. The model predicts that people will have trouble distinguishing telicity in the Endpoint Unseen condition.

To explain how speakers select among the many different verb forms that could be used for describing the actions, we introduced the Length Variation assumption in the model where the length of the verb was related to the amount of semantic information that was salient for the model. To examine whether people are sensitive to the length of videos and vary their verbs accordingly, we created shorter versions of the videos and compared them to the original longer versions. The model predicts that more unique verb forms with multiple subverbs will be produced in the long video condition than in the short video condition.

5.1. Participants

The participants were a sample of 80 Japanese people from the crowdworks.jp website. They were required to participate online from a computer in Japan. Participants were paid 550 yen for their participation. The work was conducted with the formal approval of the Ethics Review Committee for Experimental Research with Human Subjects at Kobe City University of Foreign Studies.

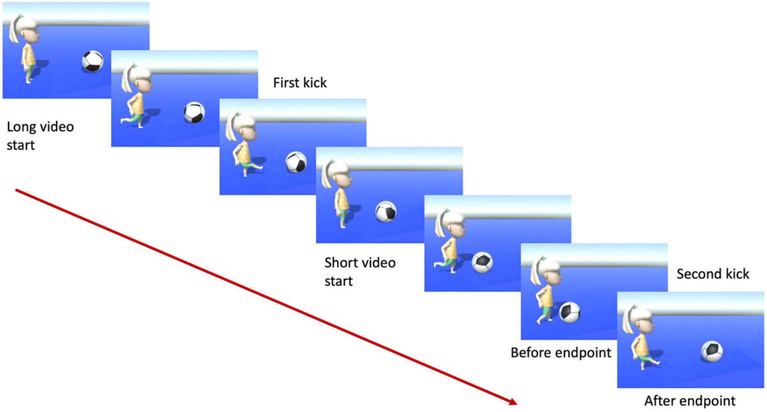


Fig. 18. Kicking action with frames used for making long/short video and before/after endpoint versions.

5.2. *Stimuli*

For each of the actions, four versions were created that crossed length (Long/Short) with endpoint (Seen/Unseen). Most of the videos were based on the previous telic videos from Experiments 1 and 2 (Fig. 18 shows an example with the kicking video). In many of the videos, the action was performed twice (e.g., the agent kicks the ball once and does not connect, so she kicks it again and the ball flies away). For each action, the endpoint was identified (e.g., the second kick which connects with the ball). The Long Unseen video started at the beginning of the original video (Long video start) and extended until right before the endpoint (Before endpoint). The Long Seen video was similar, but ended after the foot made contact with the ball and the ball moved away (After endpoint). The two Long videos were equated in length, so the Seen video started a few frames after the start of the Unseen video. In the Short versions, the videos started after the first kick (Short video start) and they were also equated in length. For videos that did not depict multiple actions (e.g., climbing), the Short video was made at least half as long as the Long video.

Our stimuli included iterative actions (Burn, Chop, Stack, and Line Up) where all possible actions were completed in the telic condition and there were additional possible goals in the atelic condition. To explain how telicity is recognized in these scenes, we used the Adaptive Tracking assumption, where additional objects become available to track once the initial objects are processed. To test this assumption visually, we used the atelic scenes to create new animations where the camera was positioned such that the additional possible goals could not be seen (Fig. 19 shows a chop scene where only two gates are available for chopping). The Seen versions would end with the final action on the visible object (Second chop completed). The Unseen version would have the same sequence, but would pan to show additional objects (two additional gates). This visually simulates the idea that object-tracking resources become available to track additional objects at the end of the scene in this condition. Long and Short versions were created for both Seen and Unseen videos.

Since the study was carried out with online participants, it was necessary to test the participants to ensure that they were native speakers of Japanese. This was done by giving them

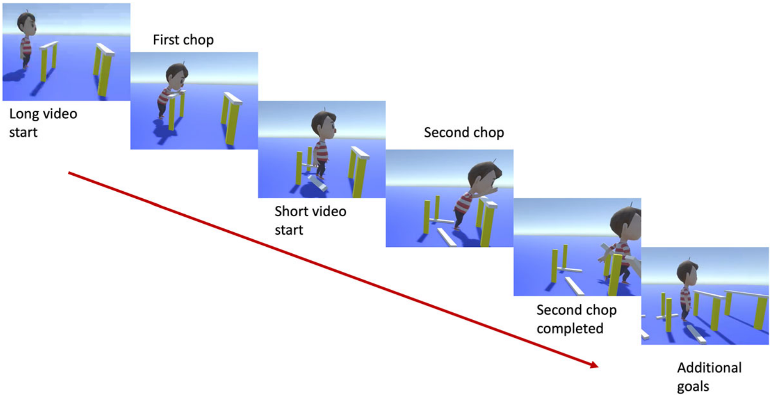


Fig. 19. Chopping action with frames used for making long/short video and before/after endpoint versions. Before endpoint ends with second chop, while after endpoint shows additional goals.

32 Japanese culture and language multiple-choice questions (accuracy on these questions was 93%). Participants saw each of the 16 actions twice with different characters, endpoints, and lengths in the two presentations. Video descriptions alternated with the culture and language questions. Four lists were created that counterbalanced telicity and length, and the order of the actions was reversed in half of the lists. There were two practice items before the main stimulus items were presented.

5.3. *Procedure*

The video trials started with a screen which reminded participants that the video was starting soon. Then the video would start automatically. Once the video ended, participants would see an instruction prompt and a text box for typing in their descriptions. When they were done, they would press the next button to go to the multiple-choice question. The question was written at the top of the screen and four answer buttons were present below. After they responded, the next video trial would begin.

Participant responses were coded using the same coding system as was used for the previous studies. There were 1227 simple past forms, 825 progressive nonpast forms, 2 progressive past forms, 492 simple nonpast forms, 0 dialect forms, and 3 other forms (e.g., fragments, errors). The analysis focused on simple past and progressive nonpast forms. As before, a second variation coding was performed to examine the use of subverbs and to test the prediction concerning the influence of length on unique verb forms. For this coding, we removed the past/progressive morphology from verbs. For each of these forms, we computed their proportion in the past tense and their log frequency in the responses.

5.4. *Results*

The first analysis examined whether endpoint and length influenced past production (Fig. 20). A logistic mixed-effects model was fitted to past production with centered video

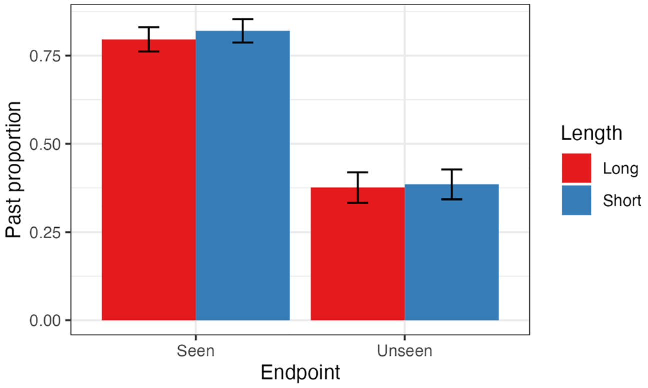


Fig. 20. Past form proportion of adult productions in Experiment 3 split by endpoint visibility and length.

endpoint and length crossed. Random effects of participant and video on the intercept were included and the maximal model that converged for the data only included random slopes for endpoint under video (fixed effects $R^2 = .242$; whole model $R^2 = .77$). There was a significant effect of endpoint visibility, where videos with endpoints that were seen were 36% more likely to be described in past tense compared to those that were unseen (excluding nonpast forms, $\beta = 0.31$, SE $= 0.27$, $\chi^2 = 19.89$, $p<.001$). There was no main effect of length and no interaction with endpoint. This demonstrated that even when a goal was present in the scene, actually seeing the endpoint increased the use of the past tense. On the other hand, the video length did not appear to influence the selection of tense/aspect morphology.

As in the previous studies, participants generated many verb forms in their descriptions. Fig. 21 shows the past proportion by rank for the verb forms. In total, there were 446 unique forms (more than 27 per action), and more than half of the verbs are used either in past only (221 verbs) or progressive only (146 verbs) compared to 79 verbs that occurred with both past and progressive endings. Even considering just verbs that occur more than once (blue diamonds), there were many verbs that occur only in one tense/aspect form. Thus, it appears that Japanese speakers are not simply selecting among a few high-frequency forms, but instead are using a large range of longer low-frequency forms with subverbs that are associated with particular tense/aspect forms.

The length manipulation in this study was designed to test whether the subverbs used were related to the amount of information in the video. The ranks in Fig. 21 are a measure of the variation in the verb forms in each action, because higher ranks only appear if the action elicited many unique verb forms as a result of particular combinations of subverbs. The Length Variation assumption predicted that long videos will elicit more unique verbs with higher ranks than short videos. To test this prediction, a mixed-effects linear regression model was fitted to average rank with centered video endpoint and length crossed (Fig. 22). Random effects of participant and video on the intercept were included and the maximal model that converged for the data included random slopes for endpoint and length under video, but none for participant (fixed effects $R^2 = .005$; whole model $R^2 = .23$). There was a significant effect

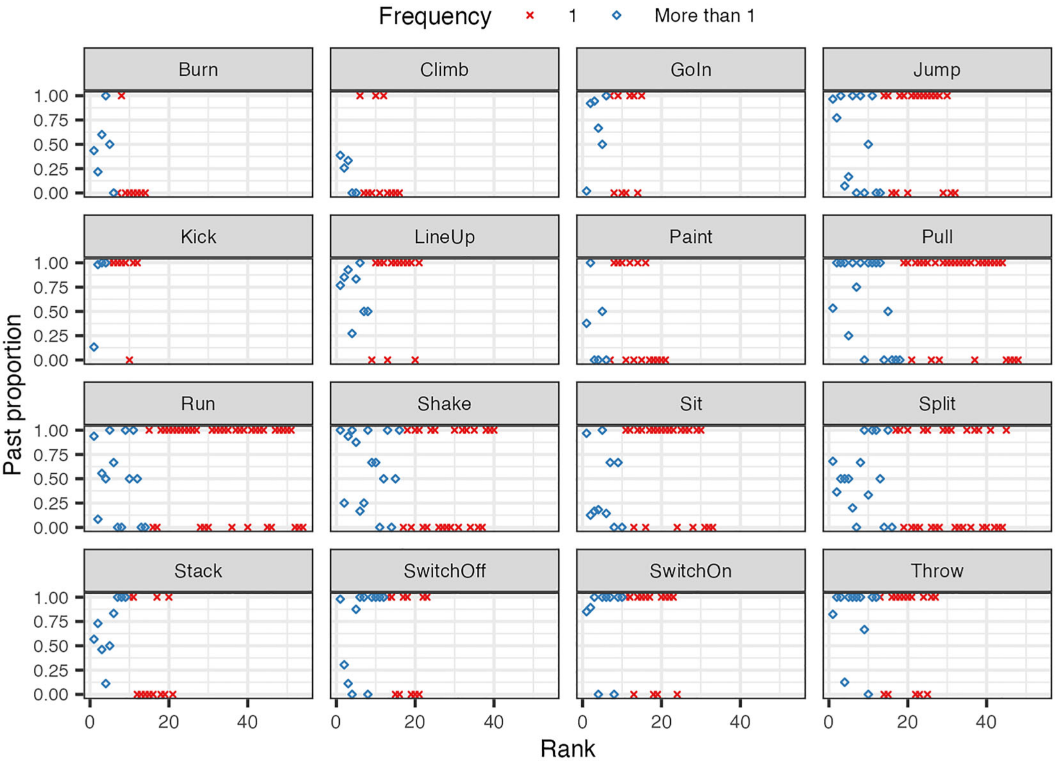


Fig. 21. Unique verb forms produced by adults in Experiment 3 for each video by past proportion and rank in our participant responses.

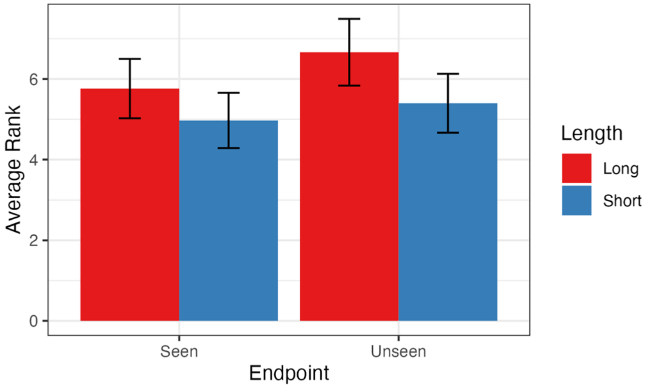


Fig. 22. Average rank of verb forms in Experiment 3 by endpoint visibility and length.

of length, where long videos yielded verb forms which had average ranks which were 1.01 higher than the short videos $\beta = -0.74$, SE $= 0.61$, $\chi^2 = 4.28$, $p = .039$. There was no main effect of endpoint and no interaction between endpoint and length. This demonstrates that longer videos increased the use of subverbs, which made the verbs more distinct from the other verbs used. However, the uniqueness of the subverbs was not strongly linked to whether the endpoint was seen or not.

6. Discussion

Research on tense and aspect in Japanese has assumed that abstract features like telicity play an important role, but it was not clear how these features are identified from visual situations. We proposed that visual cues for manner and goals could help in identifying the event types that influence tense/aspect. The first two studies examined this by presenting videos of animated actions where the manner was matched, but the goal was manipulated by varying endpoint-related object information. Both adults (Experiment 1) and children (Experiment 2) were more likely to use past tense morphology rather than progressive morphology when viewing videos with goals than videos without goals, which suggests that goal objects are used to compute the telicity of events. It was also found that speakers generated complex verbs that encoded manner and goal into different subverbs (*aruite haitta* walk + enter). These results suggest that visual cues can influence the selection of verb morphology.

To better understand how verbs might be linked to visual information, a generative deep-learning Heuristics-Verb model was developed using visual heuristics related to the manner of motion and goal-directed action. This information was fed through several LSTMs to create an action representation for the whole event and this was used as input for the Kana LSTM model, which produced a sequence of kana for verbs. Compared to the No Variation model which only used kana probabilities, the Length Variation model which had semantic information about the length of the action provided a better fit to the human distribution of verbs. As in the human data in Experiments 1 and 2, this model could vary its verb morphology for telic and atelic videos. One mismatch was the fact that the telicity distinction increased over development in the model (approximated by training epoch), but not in the human data. This may be related to the preference for progressive in children (Shirai, 1998), and future work will need to examine whether changes in the input can create a model with a closer fit to the human data.

The model's ability to distinguish telic and atelic videos was due to its sensitivity to the changes in the event at the endpoint (e.g., target knocked over by ball). This was tested by manipulating whether the video stopped before or after the endpoint. The model exhibited a telicity distinction when its input included the part of the action sequence associated with the endpoint in the telic video, but not when the endpoint was removed from its input. This prediction of the model was examined in the final experiment, and it was found that humans also use the past tense more when the video showed the endpoint. Thus, in this task, participants are parsing the visual components of actual situations and using that information to generate verb morphology.

Another important aspect of the results was the wide variation in verb forms produced by children and adults in all three studies. The Aspect Hypothesis does not discuss why speakers choose different lemmas such as the accomplishment form pull+tear (*hippatte chigit*), the achievement form pull+out (*hippari nui*), or the activity form pull+play (*hippatte ason*), but there appears to be some role of event types in the selection of the lemma. Our account proposes that both the verb lemma and subverbs, as well as the verb ending, are selected using the same visual heuristics, and therefore, events with more visual semantic information will yield verbs with additional subverbs. This claim was tested in the final experiment by changing the length of the videos, and it was found that human participants used more unique forms when shown longer videos.

The Heuristics-Verb model provides a novel model of verb production. Existing theories of word production focus on the selection of one dominant lemma (errors arise when multiple lemmas are selected, Levelt, 1989), and the event type of this main lemma has been argued to influence tense-aspect morphology (Shirai & Andersen, 1995). But in Japanese, speakers regularly produce complex verbs with multiple subverbs where identifying the component lemmas and their event types is challenging. These multiple subverbs are also not easily extracted from action-understanding systems that classify actions with a single label (Zhu et al., 2020). Furthermore, models of verb production struggle with producing low-frequency verbs (Engelmann et al., 2019), because they do not encode the additional meaning information that is used to select these verbs. Existing models of past tense production (Joanisse & Seidenberg, 1999; Plunkett & Juola, 1999) are unable to explain the way that visual components of events such as manner and goal/results influence verb morphology. The present Heuristics-Verb model addresses all of these issues by assuming that a complex multiple-component visual representation can be involved in verb production. The action representation contains manner and goal/result components that can be flexibly used to select verb lemmas, subverbs, or tense/aspect morphology.

One of the main limitations of the Heuristics-Verb model is that it is trained on a small dataset. Other action-understanding approaches work across large sets of real-world videos using a variety of human-labeled data and complex computational architectures (Zhu et al., 2020). While these models have better coverage, it is difficult to derive predictions from them about human behavior, because they use opaque representations that are not supported by human perception research and they have access to linguistic codings that may not be available to humans. In contrast, the Heuristics-Verb model used a small set of inputs that are motivated by human experimental work on language or vision (e.g., limitations on object tracking, biological motion recognition). Its ability to generalize outside of its training regime was limited. But since its inputs abstracted away from the surface visual characteristics of the scene (e.g., color, view angle), it can generalize to actions involving novel objects. Future work will need to expand the input to develop a more robust and general account of verb production.

The minimization of real-world situations in linguistic approaches arises out of the assumption of modularity between language and vision (Fodor, 1983). In contrast, the present approach argues that visual processing provides a rich set of features that can be used to explain a range of linguistic behaviors. In particular, we have shown for the first time that visual heuristics for manner and goals can be used to select lemmas and subverbs as well as tense/aspect morphology in a large set of Japanese verb forms. Given that language interacted with visually mediated situational meaning throughout its evolution, we think it is likely that visual heuristics play a role in other aspects of language acquisition and use.

Acknowledgments

This research was supported by an ESRC International Networking grant. Thanks to Nathan Elsdon for his assistance with making the videos. Franklin Chang was also supported by KAKEN grant 19K12733 from the Japanese Society for the Promotion of Science. We would like to thank the children, teachers, and Director Yasuyuki Matsuda at Kobe Gakuen-toshi YMCA Kindergarten for their help with this study.

References

Allen, S., Özyürek, A., Kita, S., Brown, A., Furman, R., Ishizuka, T., & Fujii, M. (2007). Language-specific and universal influences in children's syntactic packaging of Manner and Path: A comparison of English, Japanese, and Turkish. *Cognition*, *102*(1), 16–48. https://doi.org/10.1016/j.cognition.2005.12.006

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, (02), 239–273.

Andersen, R. W., & Shirai, Y. (1994). Discourse motivations for some cognitive acquisition principles. *Studies in Second Language Acquisition*, *16*, (2), 133–156. https://doi.org/10.1017/S0272263100012845

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, (3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology*, *41*, (2), 328–337.

Bidet-Ildei, C., Kitromilides, E., Orliaguet, J.-P., Pavlova, M., & Gentaz, E. (2014). Preference for point-light human biological motion in newborns: Contribution of translational displacement. *Developmental Psychology*, *50*, (1), 113–120.

Chan, E., Baumann, O., Bellgrove, M., & Mattingley, J. (2012). From objects to landmarks: The function of visual location information in spatial navigation. *Frontiers in Psychology*, *3*, 304.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, (5), 609–651.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, (2), 234–272.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, (2), 179–211.

Emmendorfer, A. K., Correia, J. M., Jansma, B. M., Kotz, S. A., & Bonte, M. (2020). ERP mismatch response to phonological and temporal regularities in speech. *Scientific Reports*, *10*, (1), 9917. https://doi.org/10.1038/s41598-020-66824-x

Engelmann, F., Granlund, S., Kolak, J., Szreder, M., Ambridge, B., Pine, J., Theakston, A., & Lieven, E. (2019). How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages. *Cognitive Psychology*, *110*, 30–69. https://doi.org/10.1016/j.cogpsych.2019.02.001

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, (2), 165–193.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (Vol. *9*, pp. 249–256). Sardinia, Italy: PMLR.

Golinkoff, R. M., Chung, H. L., Hirsh-Pasek, K., Liu, J., Bertenthal, B. I., Brand, R., Maguire, M., & Hennon, E. (2002). Young children can extend motion verbs to point-light displays. *Developmental Psychology*, *38*, (4), 604–614. https://doi.org/10.1037//0012-1649.38.4.604

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, (8), 1735–1780.

Jessop, A., & Chang, F. (2020). Thematic role information is maintained in the visual object-tracking system. *Quarterly Journal of Experimental Psychology*, *73*, (1), 146–163. https://doi.org/10.1177/1747021819882842

Jessop, A., & Chang, F. (2022). Thematic role tracking difficulties across multiple visual events influences role use in language production. *Visual Cognition*, *30*, (3), 151–173. https://doi.org/10.1080/13506285.2021.2013374

Ji, Y., & Papafragou, A. (2022). Boundedness in event cognition: Viewers spontaneously represent the temporal texture of events. *Journal of Memory and Language*, *127*, 104353. https://doi.org/10.1016/j.jml.2022.104353

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences*, *96*, (13), 7592–7597. https://doi.org/10.1073/pnas.96.13.7592

Jones, J. M., Vinson, D., Clostre, N., Zhu, A. L., Santiago, J., & Vigliocco, G. (2014). The Bouba effect: Sound-shape iconicity in iterated and implicit learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 2459–2464.

Kazanina, N., & Phillips, C. (2007). A developmental perspective on the Imperfective Paradox. *Cognition*, *105*, (1), 65–102.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. *arXiv:1412.6980* [Cs]. Retrieved from https://arxiv.org/abs/1412.6980

Learmonth, A. E., Newcombe, N. S., & Huttenlocher, J. (2001). Toddlers' use of metric information and landmarks to reorient. *Journal of Experimental Child Psychology*, *80*, (3), 225–244. https://doi.org/10.1006/jecp.2001.2635

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, (7553), 436–444. https://doi.org/10.1038/nature14539

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levin, B., & Hovav, M. R. (1991). Wiping the slate clean: A lexical semantic exploration. *Cognition*, *41*, (1–3), 123–151.

Maguire, M., Hirsh-Pasek, K., Golinkoff, R., Imai, M., Haryu, E., Vanegas, S., Okada, H., Pulverman, R., & Sanchez-Davis, B. (2010). A developmental shift from similar to language-specific strategies in verb acquisition: A comparison of English, Spanish, and Japanese. *Cognition*, *114*, (3), 299–319.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*, (3), 243–282.

Mathis, A., & Papafragou, A. (2022). Agents' goals affect construal of event endpoints. *Journal of Memory and Language*, *127*, 104373. https://doi.org/10.1016/j.jml.2022.104373

Matsumoto, Y. (1996). *Complex predicates in Japanese: A syntactic and semantic study of the notion 'word'*. Tokyo & Stanford: Kuroshio Shuppan & CSLI.

McShane, J., & Whittaker, S. (1988). The encoding of tense and aspect by three- to five-year-old children. *Journal of Experimental Child Psychology*, *45*, (1), 52–70. https://doi.org/10.1016/0022-0965(88)90050-1

Papafragou, A., Massey, C., & Gleitman, L. (2006). When English proposes what Greek presupposes: The cross-linguistic encoding of motion events. *Cognition*, *98*, (3), B75–B87.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*, (3), 674–685.

Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, *23*, (4), 463–490.

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, *73*, (1), 28–38. https://doi.org/10.1037/h0024135

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism*. *Spatial Vision*, *3*, (3), 179–197.

Reas, C., & Fry, B. (2006). Processing: Programming for the media arts. *AI & Society*, *20*, (4), 526–538.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*, (4), 613–643.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, (6088), 533–536.

Shirai, Y. (1991). *Primacy of aspect in language acquisition: Simplified input and prototype* [PhD thesis]. Los Angeles, CA: University of California.

Shirai, Y. (1998). The emergence of tense-aspect morphology in Japanese: Universal predisposition? *First Language*, *18*, (54), 281–309.

Shirai, Y. (2000). The semantics of the Japanese imperfective-teiru: An integrative approach. *Journal of Pragmatics*, *32*, (3), 327–361.

Shirai, Y., & Andersen, R. W. (1995). The acquisition of tense-aspect morphology: A prototype account. *Language*, *71*, (4), 743–762. https://doi.org/10.2307/415743

Shirai, Y., & Kurono, A. (1998). The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning*, *48*, (2), 279–244.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, (3), 153–181. https://doi.org/10.1037/h0046162

Supercyan. (2018). *Character pack: Common people*. Retrieved from https://assetstore.unity.com/packages/3d/characters/humanoids/character-pack-common-people-65722

Talmy, L. (1975). Semantics and syntax of motion. In J. P. Kimball (Ed.), *Syntax and semantics* (Vol. *4*, pp. 181–238). Leiden, The Netherlands: Brill.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (Vol. *3*, pp. 36–149). Cambridge University Press.

Tanimura, S., Takahashi, H., Baba, H., & Nokubi, T. (2018). Nippon R package: Japanese Utility Functions and Data.

Tatsumi, T., Ambridge, B., & Pine, J. M. (2018). Testing an input-based account of children's errors with inflectional morphology: An elicited production study of Japanese. *Journal of Child Language*, *45*, (5), 1144–1173. https://doi.org/10.1017/S0305000918000107

Tatsumi, T., Chang, F., & Pine, J. M. (2021). Exploring the acquisition of verb inflections in Japanese: A probabilistic analysis of seven adult-child corpora. *First Language*, *41*, (1), 41–66. https://doi.org/10.1177/0142723720926320

Tatsumi, T., & Pine, J. M. (2016). Comparing generativist and constructivist accounts of the use of the past tense form in early child Japanese. *Journal of Child Language*, *43*, (6), 1365–1384. https://doi.org/10.1017/S0305000915000732

Twomey, K. E., Chang, F., & Ambridge, B. (2016). Lexical distributional cues, but not situational cues, are readily used to learn abstract locative verb-structure associations. *Cognition*, *153*, 124–139. https://doi.org/10.1016/j.cognition.2016.05.001

Unity Technologies. (2018). Unity (Version 2018.4.12). Retrieved from http://unity3d.com

van Hout, A. (2016). Lexical and grammatical aspect. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford handbook of developmental linguistics* (pp. 587–610). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199601264.013.25

Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, NY: Cornell University Press.

Von Stutterheim, C., Andermann, M., Carroll, M., Flecken, M., & Schmiedtová, B. (2012). How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data, and memory performance. *Linguistics*, *50*, (4), 833–867.

Wagner, L., Swensen, L. D., & Naigles, L. R. (2009). Children's early productivity with verbal morphology. *Cognitive Development*, *24*, (3), 223–239. https://doi.org/10.1016/j.cogdev.2009.05.001

Waller, D., Loomis, J. M., Golledge, R. G., & Beall, A. C. (2000). Place learning in humans: The role of distance and direction information. *Spatial Cognition and Computation*, *2*, 333–354.

Wang, X., Chen, D., Yang, T., Hu, B., & Zhang, J. (2016). Action recognition based on object tracking and dense trajectories. In *2016 IEEE International Conference on Automatica (ICA-ACCA)*. https://doi.org/10.1109/ICA-ACCA.2016.7778391

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, (1), 1–34.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., & Shah, M. (2022). Deep learning-based human pose estimation: A survey. arXiv. https://doi.org/10.48550/arXiv.2012.13392

Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., & Li, M. (2020). A comprehensive study of deep video action recognition. *arXiv:2012.06567 [Cs]*. Retrieved from https://arxiv.org/abs/2012.06567