Abstract
How do language learners avoid the production of verb argument structure overgeneralization errors (*The clown laughed the man* c.f. *The clown made the man laugh*), while retaining the ability to apply such generalizations productively when appropriate? This question has long been seen as one that is both particularly central to acquisition research and particularly challenging. Focussing on causative overgeneralization errors of this type, a previous study reported a computational model that learns, on the basis of corpus data and human-derived verb-semantic-feature ratings, to predict adults' by-verb preferences for less- versus more-transparent
causative forms (e.g., *The clown laughed the man vs The clown made the man laugh) across English, Hebrew, Hindi, Japanese and K'iche Mayan. Here, we tested the ability of this model to explain binary grammaticality judgment data from children aged 4:0-5:0, and elicited-production data from children aged 4:0-5:0 and 5:6-6:6 (N=48 per language). In general, the model successfully simulated both children's judgment and production data, with correlations of r=0.5-0.6 and r=0.75-0.85, respectively, and also generalized to unseen verbs. Importantly, learners of all five languages showed some evidence of making the types of overgeneralization errors – in both judgments and production – previously observed in naturalistic studies of English (e.g., *I'm dancing it). Together with previous findings, the present study demonstrates that a simple discriminative learning model can explain (a) adults' continuous judgment data, (b) children's binary judgment data and (c) children's production data (with no training of these datasets), and therefore constitutes a plausible mechanistic account of the retreat from overgeneralization.

**Keywords**

child language acquisition, verb semantics, causative, English, Japanese, Hindi, Hebrew, K'iche', discriminative learning

---

This article is included in the Excellent Science gateway.

---

**Corresponding author:** Ben Ambridge (Ben.Ambridge@Liverpool.ac.uk)

**Author roles:** Ambridge B: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Doherty L: Investigation, Resources, Writing – Review & Editing; Maitreyee R: Formal Analysis, Investigation, Software, Writing – Review & Editing; Tatsumi T: Investigation, Resources, Writing – Review & Editing; Zicherman S: Investigation, Resources, Writing – Review & Editing; Mateo Pedro P: Methodology, Writing – Review & Editing; Kawakami A: Investigation, Writing – Review & Editing; Bidgood A: Methodology, Writing – Review & Editing; Narasimhan B: Methodology, Writing – Review & Editing; Arnon I: Methodology, Writing – Review & Editing; Bekman D: Investigation, Writing – Review & Editing; Fabiola Can Pixabaj S: Investigation, Writing – Review & Editing; Marroquín Pelíz M: Investigation, Writing – Review & Editing; Julajuj Mendoza M: Investigation, Writing – Review & Editing; Samanta S: Formal Analysis, Software; Campbell S: Formal Analysis, Software; McCauley S: Formal Analysis, Software; Berman R: Methodology, Writing – Review & Editing; Misra Sharma D: Methodology; Bhaya Nair R: Methodology; Fukumura K: Methodology, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No [681295]), (project CLASS). Ben Ambridge is Professor in the International Centre for Language and Communicative Development (LuCiD) at The University of Liverpool. The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2021 Ambridge B et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Ambridge B, Doherty L, Maitreyee R et al. Testing a computational model of causative overgeneralizations: Child judgment and production data from English, Hebrew, Hindi, Japanese and K'iche' [version 1; peer review: 3 approved with reservations] Open Research Europe 2021, 1:1 https://doi.org/10.12688/openreseurope.13008.1

**First published:** 24 Mar 2021, 1:1 https://doi.org/10.12688/openreseurope.13008.1
Plain language summary
When learning their native language, children often produce errors in which they use verbs in “ungrammatical” sentence types (e.g., “The clown laughed the man”, whereas an adult would say “The clown made the man laugh”). Although these examples are from English, similar errors are observed in many other languages including Hebrew, Hindi, Japanese and K’iche Mayan. A previous study reported a computer model which, when trained on an approximation of real language input, simulated the relative grammatical acceptability of these errors with different verbs as judged by child and adult raters. The aim of this study was to investigate whether the same model could explain (a) binary judgments from younger children (4–5 year-olds, who were simply asked “Is this sentence acceptable” rather than “How acceptable is this sentence?” and (b) the rates at which children learning these five languages actually produce such errors for different verbs (e.g., Someone laughed/danced/sang the boy). In general, the model performed very well on both tasks for all five languages.

Introduction
The question of how language learners come to avoid verb argument structure overgeneralization errors such as “The clown laughed the man” – in some cases after a protracted period of producing them – has been described as a “learnability paradox” (Pinker, 1989: 415); “one of the most…difficult challenges for all students of language acquisition” (Bowerman, 1988: 73). The problem is this: On the one hand, children need to be able to use verbs in argument structure constructions in which they have not witnessed them; this type of productivity is the hallmark of human language. On the other hand, children need to be able to constrain this generalization process in order to avoid producing ungrammatical utterances such as *The clown laughed the man. These types of errors, in which English-speaking children incorrectly mark causation using the transitive causative for verbs that prefer the periphrastic causative (e.g., The clown made the man laugh) are the focus of the present study; along with equivalent errors in Hebrew, Hindi, Japanese and K’iche Mayan. Further naturalistically obtained examples of this error are summarized in Table 1 below (from the diary study of Ambridge & Ambridge, 2020). Similar errors have been observed in naturalistic data for Japanese (Nakaishi, 2016; see also the experimental study of Fukuda & Fukuda, 2001), though they have not, to our knowledge, been investigated for any of the other languages included here.

<table>
<thead>
<tr>
<th>Age</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2;3</td>
<td>Can you reach me? (Already being held, wants lifting up higher to touch sparkly part of a sign)</td>
</tr>
<tr>
<td>2;4</td>
<td>Can you jump me off? (wants help jumping down off the bed)</td>
</tr>
<tr>
<td>2;4</td>
<td>Did you drop the letters? (“Did you make the letters drop?” Foam letters stuck to the bathroom wall have fallen into the bath)</td>
</tr>
<tr>
<td>2;6</td>
<td>(Dad: why are you running?) It’s practising me to run like that</td>
</tr>
<tr>
<td>2;6</td>
<td>Jump me!</td>
</tr>
<tr>
<td>2;6</td>
<td>Don’t swim me</td>
</tr>
<tr>
<td>2;7</td>
<td>Run me down, jump me down (wants to run down slide)</td>
</tr>
<tr>
<td>2;7</td>
<td>Jump me</td>
</tr>
<tr>
<td>2;7</td>
<td>Drink me. drink me, Dad! (Can’t reach juice in bottom of cup and wants it tipped right back)</td>
</tr>
<tr>
<td>2;7</td>
<td>I’m just dancing it (shaking the bent-double flap of the elephant’s door in Dear Zoo, to make it dance)</td>
</tr>
<tr>
<td>2;7</td>
<td>I can dance it (book)</td>
</tr>
<tr>
<td>2;7</td>
<td>I’m dancing it</td>
</tr>
<tr>
<td>2;7</td>
<td>This is the boat - swim it!</td>
</tr>
<tr>
<td>2;7</td>
<td>Swim that aeroplane (submarine)</td>
</tr>
<tr>
<td>2;7</td>
<td>Stay your leg up there (holding dad’s leg)</td>
</tr>
<tr>
<td>Age</td>
<td>Error</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.7</td>
<td>Stop jumping them (Dad is tapping rabbits in Peter Rabbit game to make them jump)</td>
</tr>
<tr>
<td>2.7</td>
<td>drink me a bit (wants straw held up to her mouth to drink squash in bed)</td>
</tr>
<tr>
<td>2.10</td>
<td>The sheet’s slipping me</td>
</tr>
<tr>
<td>2.11</td>
<td>jump me, Dad! x5</td>
</tr>
<tr>
<td>2.11</td>
<td>I jumped my legs. I hopped my legs</td>
</tr>
<tr>
<td>3.2</td>
<td>I stand on your feet and you walk me</td>
</tr>
<tr>
<td>3.2</td>
<td>(Mum: what happens to the rubbish when it goes outside?). It gets died.</td>
</tr>
<tr>
<td>3.5</td>
<td>(Dad, playing with Shopkins: Now what are we doing?) Chloe: Going them in. (What?) Into the bathroom</td>
</tr>
<tr>
<td>3.6</td>
<td>I’m try to duck her under (pushing Aurora doll under the seat belt of Barbie car)</td>
</tr>
<tr>
<td>3.6</td>
<td>Pens are difficult to come off the paper</td>
</tr>
<tr>
<td>3.7</td>
<td>Reach me up there (wants to see toys on top shelf)</td>
</tr>
<tr>
<td>3.7</td>
<td>It will get died (die/get killed)</td>
</tr>
<tr>
<td>3.7</td>
<td>That nearly feel me like I’m nearly falling off</td>
</tr>
<tr>
<td>3.8</td>
<td>I’m going it faster (exercise bike at airport)</td>
</tr>
<tr>
<td>3.8</td>
<td>Eat it in my mouth (pez sweet that has fallen onto floor - wants Dad to pick it up and post it into her mouth)</td>
</tr>
<tr>
<td>3.8</td>
<td>Disappear them and disappear them (scooping up bubbles in the bath)</td>
</tr>
<tr>
<td>3.9</td>
<td>Your turn to dance me, Dad (i.e., swing her around by the arms)</td>
</tr>
<tr>
<td>3.10</td>
<td>Those guys died Maleficent (watching Sleeping Beauty)</td>
</tr>
<tr>
<td>3.10</td>
<td>We died (dissolved) Mummy's special soap didn’t we, Dad?</td>
</tr>
<tr>
<td>3.11</td>
<td>Jump me up there (wants putting onto the toilet seat)</td>
</tr>
<tr>
<td>3.11</td>
<td>I wanna jump her in (Ariel doll into bath)</td>
</tr>
<tr>
<td>3.11</td>
<td>It will die you; it will make you killed</td>
</tr>
<tr>
<td>4.0</td>
<td>Mermaids have got special powers; they can die baddies</td>
</tr>
<tr>
<td>4.7</td>
<td>jump me x 2</td>
</tr>
</tbody>
</table>

Boyd & Goldberg, 2011; Blything et al., 2014; Goldberg, 2011; Harmon & Kapatsinski, 2017; Hsu & Chater, 2010; Irani, 2009; Perek & Goldberg, 2017; Robenalt & Goldberg, 2015; Robenalt & Goldberg, 2016; Twomey et al., 2014; Twomey et al., 2016, including two book-length treatments (Goldberg, 2019; Pinker, 1989). However, until a single recent study, research on the retreat from overgeneralization had been conducted exclusively on English (and mainly on dative and locative constructions).

This recent study (Ambridge et al., 2020), sought to explain how speakers learn to avoid not only causative errors in English, (e.g., *The clown laughed the man*), but also equivalent errors in Hebrew, Hindi, Japanese and K’iche’ Mayan. It also adopted a novel theoretical approach: Previous studies had attempted to explain this phenomenon in terms of three – to some extent – competing theories: preemption, conservatism via entrenchedment (both statistical-learning theories) and verb semantics. Ambridge et al. (2020) sought to unify these theories by building a computational model that yields all three effects in a single learning mechanism.

The model developed by Ambridge et al. (2020) – a simple two-layer connectionist network – is trained on input-output pairs consisting of a verb (e.g., break) and a causative type (e.g., for English, either the transitive causative or the make periphrastic causative respectively), in proportion to the frequency of each in a representative input corpus (e.g., for English, the frequency of [CAUSER] [BREAK] [CAUSEE] vs [CAUSER] [MAKE] [CAUSEE] BREAK). Other corpus utterances containing the relevant verb (e.g., intransitive [ACTOR] [BREAK]) are mapped to a catch-all “Other” output node. Crucially, the input to the model consists not only of an orthogonal (one-hot) “lexical” verb representation that uniquely identifies each verb stem, but also four “semantics” units. The (continuous) activation level of
these units is set on the basis of human ratings of four semantic properties thought to be relevant to languages’ preferences for less-transparent (e.g., X broke Y) versus more-transparent (X made Y break) causative forms respectively (e.g., Shibatani & Pardeshi, 2002). These semantic ratings were obtained by showing native adult speakers of each language an animation depicting the action described by each verb (though they were not given the verb itself) and asking them to rate:

**Event-merge:** The extent to which the causing and caused event are two separate events or merge into a single event that happens at a single time and a single point in space

**Autonomy of the causee**

**Requires:** Whether the caused event requires a causer

**Directive:** Whether causation is directive (e.g., giving an order) or physical

It is important to note that the model was not given any information regarding human judgments of the grammatical acceptability of the more- and less-transparent causative forms of each verb (which would make its learning task trivially simple, and akin to a conventional statistically regression model conducted on participants’ grammaticality judgments). At test, the model was presented with each verb (N=60) and interrogated for its prediction of a causative form (e.g., for English, transitive causative vs periphrastic causative with make: *Someone laughed the boy* vs *Someone made the boy laugh*). The resulting activation level of the corresponding output units was taken as the model’s “grammaticality judgment” for that form. These judgments were then correlated against those obtained from native speakers of each language (N=48 at each of ages 5–6, 9–10 and adults).

In general, the model achieved correlations of around \( r=0.75 \) with human judgments, showing only a small decrement in performance (i.e., slightly lower correlations) when tested on verbs that had been withheld during training, using split-half validation. This finding demonstrates that the model, like human learners, eventually reaches a point at which it is able to produce the appropriate causative form for verbs that it is encountering for the first time, on the basis of their semantics. Importantly, prior to this point, the model displays an “overgeneralization” stage analogous to that shown by children (at least for English). For example, when presented with laugh, the English model initially produces the transitive causative construction (e.g., *Someone laughed the boy*) with considerably higher probability than the periphrastic causative (e.g., *Someone made the boy laugh*). After around 12 epochs of training (each consisting of 10,000 corpus utterances) the probabilities begin to flip, and the model asymptotes at predictions of around 0.7 vs 0.3 for the periphrastic- versus transitive-causative respectively (“Other” uses are around zero, since the model is interrogated for a causative form).

While these findings constitute support for the model developed by Ambridge et al. (2020), this support is currently limited, since the model was assessed only on its ability to predict grammaticality judgment data obtained from older children (5–6 and 9–10 years) and adults. However, the available English data (e.g., Ambridge & Ambridge, 2020; Pinker, 1989; Bowerman, 1988) suggest that the majority of such overgeneralization errors are produced before this age. Indeed, for languages other than English, there is no more than anecdotal evidence that children produce such errors at all (either at age 5–6 or younger).

The present study therefore has two aims. The first is to test the ability of the computational model developed by Ambridge et al. (2020) to explain grammaticality judgment data from younger children than those tested previously; children aged 4;0–5;0, which necessitates the use of a binary judgment task (rather than the Likert-scale task used with children aged 5;6–6;6). The second aim is to test the ability of this computational model to explain children’s production data, including possible overgeneralization errors, at ages 4;0–5;0 and 5;6–6;6 (for comparability with the present judgment study and that of Ambridge et al., 2020, respectively).

**Ethics statement**

For both Study 1 and Study 2, ethics approval was obtained from the University of Liverpool (approval number RETH001041), as the institution with overall responsibility for the project, and from local ethics committees at the Hebrew University of Jerusalem (22032020), the International Institute of Information Technology Hyderabad (IIITH/IEC/2016/1), and the Universidad del Valle de Guatemala (¿Cómo los niños adquirían la estructura de oraciones en K’iche’?). Japanese universities do not routinely provide ethics review for psychological or linguistic research. In lieu, we therefore obtained a review from Shunzo Majima, Associate Professor at the Center for Applied Ethics and Philosophy, Hokkaido University. Parents/caregivers gave informed written consent on behalf of their children, who provided verbal assent. Written consent included both participation in the study and inclusion of the data in an anonymized publicly-available dataset.

### Study 1: Binary grammaticality judgments (4;0–5;0)

**Methods**

**Preregistration.** The sample size, materials, data collection methods and analysis plan were pre-registered at https://osf.io/qhnjk, on 15th May 2018, before data collection began. We deviate here from our planned data analysis plan, which was designed to constitute separate tests of the preemption, entrenchment and verb semantics hypothesis. In our view, such an analysis is no longer meaningful, given that (a) Ambridge et al. (2020) reported extremely high levels of collinearity between the preemption and entrenchment predictors \( r=0.75-0.96 \) for difference scores, depending on the language) and (b) our goal is now to test the computational model of Ambridge et al. (2020) which

---

1 The periphrastic causative form is termed the more-transparent form because it includes an overt causative marker (make). More- and less-transparent causative forms for Hebrew, Hindi, Japanese and K’iche’ are set out in the Methods section.
collapses the distinction between preemption, entrenchment and verb semantics into a single learning mechanism. That said, the analyses we report are “pre-registered” in the sense that they correspond directly to those reported in the computational modeling section of Ambridge et al. (2020); the only difference being that the by-verb predictor variable averages across participants’ binary grammaticality judgments (Study 1) or binary production data (Study 2), rather than continuous grammaticality judgments. As such, other than the decision to switch to these analyses in the first place, we have retained no researcher degrees of freedom (Wicherts et al., 2016). To be explicit, we are not switching our analysis plan because the original plan failed to yield a particular pattern of results: We have not conducted the analyses specified in the original analysis plan.

Computational model. The model architecture was identical to that reported in Ambridge et al. (2020; see the present Introduction for a brief outline), though new model runs were conducted (48 runs for each of 50 epochs, for each language, as in Ambridge et al., 2020).

Participants. Our preregistered analysis plan said that we would recruit 48 children aged 4;0-5;0 for each language: English, Hebrew, Hindi, Japanese and K’iche’. We achieved this target for every language except K’iche’ (N=32), for which testing had to be terminated early due to the coronavirus pandemic. All children were native learners of the relevant language, although many would have had some limited exposure to English (particularly the Hindi-speakers) and – for K’iche’ speakers – Spanish. The target sample of N=48 per language was specified in the initial grant application, but was arrived at informally on the basis of the first author’s previous work, not a power calculation. Children were recruited via schools/nurseries in the UK, Israel, India, Japan and Guatemala. Because the full set of 120 judgments would have been too onerous for young children, each child completed 60 judgments – more- and less-transparent forms for each of 30 verbs – according to one of four counterbalance lists (which can be viewed at https://osf.io/hsm3b/). These 60 judgments were split into two sessions of 30, given either on different days or on the same day with a break in between. For each child, 16 (or 14) verbs were rated in both more- and less-transparent form in the same session; the remaining 14 (or 16) verbs were rated in more-transparent form in one session and less-transparent form in the other session. A video of the procedure can be found at https://osf.io/fqyps/.

Procedure. Data were collected between January 2018 and March 2020 in schools and nurseries in the UK, Israel, India, Japan and Guatemala. Because the full set of 120 judgments would have been too onerous for young children, each child completed 60 judgments – more- and less-transparent forms for each of 30 verbs – according to one of four counterbalance lists (which can be viewed at https://osf.io/hsm3b/). These 60 judgments were split into two sessions of 30, given either on different days or on the same day with a break in between. For each child, 16 (or 14) verbs were rated in both more- and less-transparent form in the same session; the remaining 14 (or 16) verbs were rated in more-transparent form in one session and less-transparent form in the other session. A video of the procedure can be found at https://osf.io/fqyps/.

The procedure, which involved the child placing a small animal toy on a green tick or a red cross, indicating “grammatical” and “ungrammatical”, respectively (Theakston, 2004), is best summarized by the instructions that were given to children (in translation):

We are going to play a game. This dog is trying to learn to speak English (Hindi etc.). So, we’re going to watch some short videos, and he’s going to tell us what’s happening. We have to help him by telling him when he says it right, and when he gets it wrong and says it a bit funny. In the game, we will watch a cartoon and the dog

<table>
<thead>
<tr>
<th>Table 2. Less-transparent and more-transparent causative sentences for the verb LAUGH for each language. For the more-transparent causative, the overt causative marker is shown in bold type.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Less-transparent causative</strong></td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>Hebrew</td>
</tr>
<tr>
<td>Hindi</td>
</tr>
<tr>
<td>Japanese</td>
</tr>
<tr>
<td>K’iche’</td>
</tr>
</tbody>
</table>
will tell us what happens. We have to listen to the dog and then if he says something that sounds okay we put the toy on the tick [demonstrates to child] and if he says something that sounds a bit silly then we put the toy on the cross [demonstrates to child, then completes practice trials 1 (tick) and 2 (cross). Child completes practice trials 3 (tick) and 4 (cross)]. We’re going to play the game again, but this time the cartoons are going to look a bit different [shows still of animation]. They’re going to have either this little boy or something else on this stage. These big red curtains are going to close, and you have to imagine that there is someone is behind the curtains and that person is going to do something to make something change, so that when the curtains reopen you can see how its changed. So, let’s see how this one changes. [plays example animation: dress]. So as you can see, in this cartoon the person behind the curtains has done something to help or make the boy get dressed. So, when we play the game again all the sentences our dog is going to say are going to start with someone and that is who the someone is, the person behind the curtains. But we’re going to play the game the same where we watch the cartoon, the dog says the sentence and we listen and then we put the toy on the tick if it sounds okay or the cross if it sounds a bit silly. You’ve also got this grid. To win the game you need to fill all these boxes with a sticker. You’ll get a sticker every time you hear this sound [plays dog barking sound effect]. Once there is a sticker in all of the boxes you win.

The practice trials referred to are (1) The cat drank the milk, (2) *The dog the ball played with, (3) The frog caught the fly, (4) *His teeth the man brushed (or sentences with equivalent word order errors in the other languages). The example animation with dress was created solely for use as an example, and did not appear in the main stimulus set (or in Study 2). The barking sound effect was automatically triggered by the software displaying the animations (PsychoPy 2; Peirce et al., 2019), such that the child completed her grid and won the game on the final trial of each day. The experimenter also used this software to record the child’s response for each trial (grammatical, ungrammatical, equivocal/refused to answer). Responses of the latter type, which were very rare, were discarded for all statistical analyses.

Analysis. All analyses were conducted in R (version 3.6.3; R Core Team, 2020). All computational models were built using the nnet package (version 7.3-14; Venables & Ripley, 2002). Correlations were conducted using the cor function of base R. All plots were made using ggplot2 (version 2.2.1; Wickham, 2016).

Results: Binary grammaticality judgments (4:0-5:0) Before proceeding to test the computational model, it is instructive to compare children’s binary judgment data against the gold-standard adult continuous judgment data reported by Ambridge et al. (2020) in order to determine (a) whether children aged 4:0-5:0 give meaningful judgments and (b) whether they make judgments that correspond to overgeneralization errors, rating as “acceptable” sentences that receive low acceptability ratings from adults.

These data are plotted in Figure 1–Figure 3 for less-transparent forms (e.g., *Someone laughed the boy), more-transparent forms (e.g., Someone made the boy laugh) and difference scores (less- minus more-transparent forms), respectively. The x-axis shows, for each verb form, the mean acceptability rating given by adults on the five-point scale. The y-axis shows, for each verb form, the proportion of children accepting that form (recall that each child makes only a single binary acceptability judgment for each form). Forms are colour coded to indicate child judgments that correspond to “overgeneralization errors” at the group level. This was done by converting by-verb mean adult acceptability judgments and by-verb child acceptability proportions into Z-scores, and subtracting the former from the latter (or vice-versa for the difference scores, where smaller scores correspond to overgeneralization). A large positive score (red) represents overgeneralization. For example, in Figure 3 (less-transparent forms), English dance and sing are red, since around 75% of children deemed *Someone danced the boy and *Someone sang the boy to be acceptable, despite the fact that adults assigned mean acceptability ratings close to the minimum possible (1/5) for both. A large negative score (green) represents undergeneralization. For example, in Figure 3 (less-transparent forms), English break and crush are green, since only around 30–40% of children deemed Someone broke the truck and Someone crushed the can to be acceptable. Informally, the researchers who worked with the children reported that this is probably due to children rating sentences, to some extent, on the basis of the social desirability of the events described.

Such effects – as well as any other idiosyncratic (dis)preferences for particular verbs – are washed out by the difference scores, since the less- and more-transparent forms are matched for social desirability (and for the animation illustrating the event). Inspection of these scores (Figure 3) indicates that children’s judgments in fact mirrored adults’ judgments quite closely, with relatively few clear cases of overgeneralization (corresponding here to a smaller difference score for children than adults).

In order to verify that, despite some evidence of over- and under-generalization errors, children’s judgments generally mirrored those of adults, we conducted Pearson correlations on the means for each verb, corresponding to those plotted in Figure 1–Figure 3 (see Table 3).

These data suggest that, at least for English-, Hebrew-, Hindi- and K’iche’-speaking, children were indeed giving meaningful judgments. The Japanese-speaking children, however, displayed an anomalous pattern of judgments, rating as unacceptable many forms that are highly acceptable to both experimentally-tested adults (Ambridge et al., 2020) and, informally, to the present native-Japanese-speaking co-authors. Although the majority of child participants were tested in a different area of Japan to the adults tested by Ambridge et al. (2020) (Fukuyama City,
Figure 1. Child binary judgments (present study) versus adult continuous judgments for less-transparent forms.
Figure 2. Child binary judgments (present study) versus adult continuous judgments for more-transparent forms.
Figure 3. Child binary judgments (present study) versus adult continuous judgments for difference scores (less- minus more-transparent).
Table 3. By-verb correlations between child binary grammaticality judgments (mean proportion of children accepting each form) and adult continuous grammaticality judgments (rating on five-point scale; from Ambridge et al., 2020). Significant correlations are shown in bold.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Hebrew</th>
<th>Hindi</th>
<th>Japanese</th>
<th>K’iche’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transparent</td>
<td>0.31</td>
<td>0.55</td>
<td>0.59</td>
<td>-0.16</td>
<td>0.24</td>
</tr>
<tr>
<td>More</td>
<td>0.24</td>
<td>0.57</td>
<td>0.60</td>
<td>-0.08</td>
<td>0.40</td>
</tr>
<tr>
<td>Difference scores</td>
<td>0.61</td>
<td>0.68</td>
<td>0.80</td>
<td>-0.10</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Hiroshima, as opposed to Tokyo), we are not aware of any relevant dialectal differences. A possible cause of this anomalous pattern is that children are basing their ratings on social desirability, with many of the forms that they deemed unacceptable denoting undesirable actions (e.g., the less transparent forms for cry, break, steal, dissolve, bury, throw, see Figure 1, and the more-transparent forms for shiver, frighten, cry and freeze, see Figure 2). As noted above, informally, the experimenters observed this problem to some extent across all languages. It is possible, however, that social desirability may be particularly salient in the more collectivist Japanese culture (e.g., Johnson & Van de Vijver, 2003). However, social desirability alone cannot explain why the anomalous pattern holds even for difference scores which control for such by-verb effects.

Moving on to the tests of the computational model, Figure 4–Figure 8 plot – for English, Hebrew, Hindi, Japanese and K’iche’, respectively – model-child correlations for (a) the full set of 60 verbs, and (b) the split-half validation test (30 verbs, randomly selected for each run), as well as the developmental pattern shown by the model for a number of example verbs. For children’s judgments, the dependent measure is again the proportion of children judging the particular verb form (more-/less-transparent) to be acceptable on the binary judgment task (or a less-minus-more-transparent difference score). The predictor variable is the mean activation level of the corresponding unit of the model (or a difference score calculated in the same way).

In general, the model does a good job of predicting children’s binary judgment data, though less so than for adults’ continuous judgment data (Ambridge et al., 2020, reported correlations mainly in the region of r=0.75). For the present binary judgment data, focussing on difference scores, the model achieved correlations in the region of r=0.5–r=0.6 for the English, Hebrew and Hindi child data, both for seen verbs and in the split-half validation test. All six correlations are comfortably statistically significant at p<0.01 (Critical r [df=58] value for p<0.05 = 0.21; for p<0.01 = 0.30 [one tailed]). The model fares less well at predicting the raw proportions of “acceptable” judgments for less- and more-transparent causative forms; though with r values in the region of r=0.25–r=0.5, all twelve correlations are again statistically significant.

For Japanese and K’iche’ the model achieves only one significant correlation, for more-transparent causative forms in Japanese. The poor performance of the K’iche’ model was to be expected on the basis of Ambridge et al. (2020) who found similar results for adults, which they attributed to difficulties with obtaining reliable corpus counts and semantic ratings. The poor performance of the Japanese model probably reflects the fact that – as noted above – Japanese children showed the nosiest performance on the judgment task and, indeed, no significant correlation with adult judgments (possibly due to increased social-desirability effects).

Discussion: Binary grammaticality judgments (4;0-5;0)

Data from the binary judgment task show that, with the apparent exception of Japanese, children aged 4;0-5;0 are capable of providing meaningful grammatical acceptability judgments for sentences containing more- and less-transparent causative verb forms, though they also show some evidence of judgments that correspond to overgeneralization errors (e.g., accepting “Someone danced the boy” and “Someone sang the boy”). These judgment overgeneralization errors correspond to production overgeneralization errors observed in children’s spontaneous speech data (e.g., Table 1). The computational model developed by Ambridge et al. (2020) successfully explained children’s judgment data for English, Hebrew and Hindi. Its failure to do so for K’iche’ and Japanese appears to be attributable to noise in the predictor variables and children’s judgment data respectively. These findings raise two questions: (1) Do children learning each of these languages actually produce these types of overgeneralization errors and, if so, (2) Can the computational model developed by Ambridge et al. (2020) explain their by-verb patterning?

Study 2: Elicited production (4;0-5;0 and 5;6-6;6)

Methods

Preregistration. As for Study 1, the sample size, materials, data collection methods and analysis plan were pre-registered at https://osf.io/qhnjk before data collection began. Again, we depart here from our data-analysis plan in order to test the computational model of Ambridge et al. (2020) which we judge to supersede the single-process theories tested in our original pre-registration.

Computational model. As for Study 1, the model architecture was identical to that reported in Ambridge et al. (2020) though new model runs were conducted (again, 48 runs for each of 50 epochs, for each language).

Participants. As per our preregistration, we recruited 48 children at each of ages 4;0-5;0 and 5;6-6;6 for each language (including K’iche’). Children were recruited from the same populations as Study 1, though none took part in both studies. Sample size criteria, eligibility criteria, and sources and methods of participant selection were the same as for Study 1.
Figure 4. Model-child correlations for English binary judgment data.
Figure 5. Model-child correlations for Hebrew binary judgment data.
Figure 6. Model-child correlations for Hindi binary judgment data.
Figure 7. Model-child correlations for Japanese binary judgment data.
Figure 8. Model-child correlations for K’iche’ binary judgment data.
Stimuli and materials. This study used a priming methodology, in order to encourage children to attempt to produce both less- and more-transparent causative forms for each of 60 target verbs (the same set used in Study 1 and Ambridge et al., 2020). For each language, a further 60 verbs – 30 each that prefer the more- and less-transparent causative form – were selected for use as prime verbs, and corresponding animations created (following the same format as the animations for the target verbs). Only 60 prime verb were necessary, because – as for Study 1 – each child completed only half of the total number of target trials: That is, for each of 30 verbs – according to eight counterbalance lists – children described a causal animation following priming with (a) a more-transparent causative and (b) a less-transparent causative. As for Study 1, children completed two separate sessions. For each child, 16 (or 14) of the verbs appeared following both more- and less-transparent causative primes in the same session; the remaining 16 (or 14) appeared following a more-transparent causative prime in one session and a less-transparent causative prime in the other.

Procedure. Data were collected between January 2018 and March 2020 in schools and nurseries in the UK, Israel, India, Japan and Guatemala. A video of the production priming procedure can be found at https://osf.io/hq99p/. Again, the procedure, is best summarized by the instructions that were given to children (in translation):

We are going to play a game. We’re going to watch some short videos and take it in turns telling this dog what has happened. The dog has either my card or your card: If we hear this sound [plays howl sound effect] then he has mine, if we hear this [plays bark sound effect] then he has yours. Then we can put our card on the grid and whoever fills their grid first wins the whole game. Our videos are going to look a bit like this. There is a stage like one you would see in a theatre with big red curtains [plays an example animation: dress]. So, as you can see, there was a little boy on the stage and he has no top on [shows still of the stage at the beginning] and when the curtains reopened he had a top on [shows still of the stage at the end]. You must imagine that when the curtains are closed there is someone behind the curtains [shows the closed curtains]. So, in this one there was someone behind the curtains that did something to get the boy dressed. Let’s start with some practice ones and I’ll help you:

**Practice trial 1 – (dress and wrap)**

Experimenter: “someone dressed the boy”

Experimenter: “someone wrapped the present” [encourages child to repeat]

**Practice trial 2 – (hiccup and jump)**

Experimenter: “someone made the boy hiccup”

Experimenter: “someone made the boy jump” [encourages child to repeat]

**Practice trial 3 – (free and close)**

Experimenter: “someone freed the boy” [waits for/encourages child to produce…]

Child: “Someone closed the door” [experimenter corrects if necessary]

**Practice trial 4 – (burp and drink)**

Experimenter: “someone made the boy burp” [waits for/encourages child to produce…]

Child: “someone made the boy drink” [experimenter corrects if necessary]

The child and experimenter then completed the test trials in the same way. Note that the training trials were designed to give the child practice at producing less- and more-transparent causative forms following less- and more-transparent causative primes respectively. As for Study 1, the training verbs/animations did not feature in the test trials, and the barking/howling sound effects were automatically triggered by the software displaying the animations (Processing 2; https://processing.org/), such that the child completed her grid and won the game on the final trial of each day. Children’s responses were coded as to whether they included a more-transparent or less-transparent form of the target verb, with all other responses (e.g., intrusive use of the target verb; use of a different verb; no response) treated as missing data.

Analysis. All analyses were conducted in R (version 3.6.3; R Core Team, 2020). All computational models were built using the nnet package (version 7.3-14; Venables & Ripley, 2002). Correlations were conducted using the cor function of base R. All plots were made using ggplot2 (version 2.2.1; Wickham, 2016).

Results: Elicited production (4;0-5;0 and 5;6-6;6)

As for Study 1, before proceeding to test the computational model, it is instructive to compare children’s data against the gold-standard adult continuous judgment data reported by Ambridge et al. (2020) in order to determine (a) whether children’s productions generally seem to follow the constraints of the adult grammar and (b) whether they nevertheless produce overgeneralization errors that correspond to those observed (for English) in naturalistic data.

These data are plotted in Figure 9 (children aged 4:0-5:0) and Figure 10 (children aged 5:6-6:6). The x-axis shows, for each verb form, adults’ mean difference score (preference for less-over more-transparent causative forms). The y-axis shows the proportion of trials on which children, as a group, produced the less- versus more-transparent causative form of each verb (recall that all other responses were discarded as missing data).

Overgeneralization errors, this time in production, are colour coded in the same way as for Study 1. Learners of all five languages show evidence of making overgeneralization errors at relatively high rates, almost exclusively by producing less-transparent causative forms for verbs that strongly prefer
Figure 9. Children's (4;0-5;0) elicited productions (present study) versus adult continuous judgments.
Figure 10. Children’s (5;6-6;6) elicited productions (present study) versus adult continuous judgments.
more-transparent causative forms. This asymmetry is also observed for English naturalistic data (see Table 1; Bowerman, 1988) and is simulated by the computational model reported in Ambridge et al. (2020). The cause for the model (and, presumably, children) is that less-transparent causative forms are far more frequent in children’s input. For example, English-speaking 4–5 year-olds produced *Someone barked the dog and *Someone sang / crawled / wrote / whispered / sang / slept / sat the boy (c.f., Someone made the boy dog bark / sing / crawl etc.) at rates of 10–30%. Hebrew-speaking 4–5-year-olds produced corresponding errors for dissolve, turn, freeze, frighten, dance, shiver and come at rates of 40–80%, perhaps reflecting the fact that the Hebrew binyan system is, in general, relatively productive. Hindi- and Japanese-speaking 4–5-year-olds produced considerably fewer errors of this type, though still a handful (e.g., for crawl, sing; speak, whisper and hide). K’iche-speaking 4–5 year olds produced very high rates of apparent overgeneralization errors, but note from Figure 9 that the K’iche’ speaking adults show much smaller difference scores than adult speakers of the other languages. That is, while K’iche-speaking 4–5-year-olds produce less-transparent causative forms of come, speak, play, look and float more often than would be expected on the basis of adult grammatical acceptability judgments, these same judgments suggest that these forms are not strongly unacceptable.

Comparison of Figure 9 (4;0-5;0) and Figure 10 (5;6-6;6) indicates that, by this later age, overgeneralization errors have all but ceased for English, Hindi and Japanese, and decreased considerably for Hebrew. Only for K’iche’ do rates remain high, probably reflecting the fact that the dispreferred forms are not in fact deemed highly unacceptable by adults. Importantly, the productions of both the younger and older groups show significant correlations with adult rating data (see Table 4), suggesting that children understand the task and, despite the presence of some overgeneralization errors, are largely producing appropriate responses (Critical $r$ [df = 58] value for $p < 0.05 = 0.21$; for $p < 0.01 = 0.30$ [one tailed]).

Moving on to the tests of the computational model, Figure 11 plots – for English, Hebrew, Hindi, Japanese and K’iche’ respectively – model-child correlations for (a) the full set of 60 verbs, and (b) the split-half validation test (30 verbs, randomly selected for each run), as well as the developmental pattern shown by the model for a number of example verbs. Separate correlations are run for less-transparent and more-transparent causative forms because, although these sum to 1 for children (since all other responses are treated as missing data), the same is not true for the model which has three output units, corresponding to less-transparent, more-transparent and “Other”. That said, since the model rapidly learns to predict “Other” forms with very low probability when interrogated for a causative form, the correlations for less- and more-transparent forms are extremely similar.

For all languages except K’iche’, the model does an excellent job of predicting children’s judgment data with correlations upwards of $r=0.75$ for seen verbs, and $r=0.5$ for unseen verbs. Again, its poor performance with K’iche’ is likely attributable to difficulties with obtaining reliable corpus counts and semantic ratings (Ambridge et al., 2020). For this reason, we did not proceed to the split-half validation test for K’iche’. For the four other languages, however, the model’s ability at predicting children’s production data is on a par with its ability at predicting adults’ continuous judgment data (Ambridge et al., 2020). The only notable shortcoming of the model is that although it simulates the overall generalization-then-retreat pattern shown by children (see Figure 4–Figure 8, bottom panels), it does not simulate the observed differences between the present 4;0-5;0 and 5;6-6;6 year olds (see Figure 9–Figure 10). That is, the model does not show an “immature” stage in which its predictions correspond more closely to the productions of the younger than the older children. This may be because the main difference between 4;0-5;0 and 5;6-6;6 year olds is simply an across-the-board decrease in the production of overgeneralization errors, rather than any change in their by-verb patterning. Indeed, other than for forms that show floor or ceiling effects (100% or 0% less- vs more-transparent forms), an across-the-board decrease in errors that applied equally to all verbs would not affect the magnitude of the correlation.

### Discussion: Elicited production (4;0-5;0 and 5;6-6;6)

Data from the elicited-production task show that, with the exception of K’iche’, children aged 4;0-5;0 and 5;6-6;6 not only produce causative overgeneralization errors (*Someone sang / crawled / wrote / whispered / sang / slept / sat the boy; c.f., Someone made the boy dog bark / sing / crawl etc.) but do so in such a way that their by-verb patterning is well predicted by the computational model of Ambridge et al. (2020).

### General discussion

The question of how language learners (eventually) come to avoid the production of verb argument structure overgeneralization errors (*The clown laughed the man) has long been seen as one that is both particularly central to acquisition research and particularly challenging (Bowerman, 1988; Pinker, 1989). Focussing on causative overgeneralization errors of this type, Ambridge et al. (2020) built a computational model that learns, on the basis of corpus data and human-derived verb-semantic-feature ratings, to predict adults’ by-verb preferences for
Figure 11. Model-child correlations for elicited production data.
less- versus more-transparent causative forms (e.g., *The clown laughed the man vs The clown made the man laugh) across English, Hebrew, Hindi, Japanese and – to a lesser extent – K’iche. The aim of the present study was to investigate whether children learning these languages indeed produce such errors, and rate them as acceptable in a binary judgment task, and – if so – whether the computational model of Ambridge et al. (2020) can explain their patterning.

In general, the answer to these questions is a resounding “yes”. For example, the English sentences *Someone danced the boy and *Someone sang the boy were deemed acceptable by a majority of children aged 4:0-5:0 in a binary judgment task (Study 1), and were even produced at rates of around 5% and 15% respectively by (different) children at this age, though not by children aged 5:0-5:6 (Study 2). The computational model developed by Ambridge et al. was able to predict the by-verb patterning of both children’s binary-judgment data (correlations in the region of r=0.5-0.6) and their elicited-production data (correlations upwards of r=0.75), as well as generalizing to unseen verbs in a split-half validation. Given that an identical model can predict (a) adults’ continuous judgment data, (b) children’s binary judgment data and (c) children’s production data – without having been trained on any of these datasets – the problem of how language learners come to appropriately constrain their argument structure generalization looks close to being solved.

A number of issues, however, do remain. First, despite its overall successes, the model did not significantly predict Japanese children’s binary grammaticality judgments or any of the K’iche’ data (for adults and children alike). While it is possible to come up with an apparently-reasonable explanation in each case, future work should investigate the alternative possibility that the computational model tested here perhaps does not apply universally. For Japanese binary judgments, the model’s failure is almost certainly due to a task effect, since the model does successfully predict both adults’ continuous judgments and children’s production data. For K’iche’ it is less clear. Although, as already noted, both the corpus and semantic-rating data are questionable, we should not discount the possibility that this model – and the account of causatives that it instantiates – is not well suited to languages like K’iche’ that have both transitivizing and intransitivizing morphological processes. For example, in English, Hebrew and Japanese, laugh is perhaps the single most prototypical example of a highly intransitive verb that strongly prefers the less-direct, more transparent causative (e.g., Someone made the boy laugh > *Someone laughed the boy). Yet in K’iche’, intransitive laugh is derived from the transitive (though not transitive-causative) verb laugh at, and is – broadly speaking – acceptable in both causative forms; the same is true for look (derived from look at) and speak (from speak about). Perhaps, then, the crosslinguistic typology of causatives embodied by the computational model tested here is not quite accurate.

This relates to a second issue: While it is certainly impressive that the model can account for adult and child data across – K’iche’ aside – four unrelated languages; these four languages hardly constitute a large or representative sample of all the languages of the world. Future work using the methods here should investigate whether this model generalizes to other languages.

Third, future work using related methods should investigate whether an account of this type can explain the retreat from overgeneralization for a wide variety of syntactic and morphological constructions. We see no particular reason to believe that it cannot (e.g., see Ambridge & Blything, 2016; Li & MacWhinney, 1996, for similar models of the English un-prefixation and dative constructions), but, of course, the outcomes of such investigations cannot be anticipated.

Fourth, even for the restricted case of less- versus more-transparent causative forms, the model tested here does not solve the learning problem entirely, given that it starts from the point at which children have already acquired the relevant forms (e.g., the transitive-causative and make periphrastic causatives for English; lexical causatives and the –(s)ase causative marker for Japanese; the transitive and causative binyanim for Hebrew). Although the model learns a great deal about the meanings of these forms – i.e., the particular type of causation that is associated with each – the forms themselves are pre-given; and in most cases are highly abstract generalizations. In this respect, the account tested here is no different to all other accounts of the retreat from overgeneralization discussed in the Introduction. But until we have a model that can learn the generalizations in the first place, we cannot quite say that the problem of forming appropriately restricted generalizations has been solved.

Finally, the present study has important methodological implications in that three different methods – continuous grammaticality judgments, binary grammaticality judgments and elicited production – have produced findings that are generally very highly correlated with one another. Indeed, we could – at a push – argue that five different methods have converged on similar conclusions, if we include both the diary data that first uncovered such errors (e.g., Bowerman, 1988; Table 1) and the corpus analysis used to derive the model’s training data. The methodological implications are – on the one hand – that triangulating different methods on the same set of stimuli provides a particularly detailed and robust test of a particular model; and – on the other – that where this is not possible, we can be reasonably confident that conclusions drawn on the basis of data collected using one method will generalize to another.

In conclusion, while work remains to be done to extend this research to other constructions and other language families, the present findings that the computational model developed by Ambridge et al. (2020) explains both children’s binary grammaticality judgment and elicited production data across a range of languages suggest that a solution to the longstanding problem of the retreat from overgeneralization is within our grasp.

**Data availability**

**Underlying data**

This project contains the following underlying data:

**Binary Judgments And Production Zip** (Zip file containing each of the following)

**Binary Modeling (Folder containing each of the following)**
- Binary Modeling.R (R code for the computational modeling)
- ENG_Adults.csv – English grammaticality judgment data (from Ambridge et al., 2020)
- ENG_Input.csv – English input file for the computational modeling
- ENG_Results.csv – English children’s binary judgment data – target for modeling
- HEB_Adults.csv – Hebrew grammaticality judgment data (from Ambridge et al., 2020)
- HEB_Input.csv – Hebrew input file for the computational modeling
- HEB_Results.csv – Hebrew children’s binary judgment data – target for modeling
- HIN_Adults.csv – Hindi grammaticality judgment data (from Ambridge et al., 2020)
- HIN_Input.csv – Hindi input file for the computational modeling
- HIN_Results.csv – Hindi children’s binary judgment data – target for modeling
- JAP_Adults.csv – Japanese grammaticality judgment data (from Ambridge et al., 2020)
- JAP_Input.csv – Japanese input file for the computational modeling
- JAP_Results.csv – Japanese children’s binary judgment data – target for modeling
- KIC_Adults.csv – Kiche’ grammaticality judgment data (from Ambridge et al., 2020)
- KIC_Input.csv – Kiche’ input file for the computational modeling
- KIC_Results.csv – Kiche’ children’s binary judgment data – target for modeling

**Production Modeling (Folder containing each of the following)**
- ENG_Input.csv – English input file for the computational modeling
- ENG_Results.csv – English children’s production data – target for modeling
- HEB_Input.csv – Hebrew input file for the computational modeling
- HEB_Results.csv – Hebrew children’s production data – target for modeling
- JAP_Input.csv – Japanese input file for the computational modeling
- JAP_Results.csv – Japanese children’s production data – target for modeling
- KIC_Input.csv – Kiche’ input file for the computational modeling
- KIC_Results.csv – Kiche’ children’s production data – target for modeling
- Production Correlations with Old Paper.R (R files for creating Figures 1–3)

**Extended data**


This project contains the following extended data:
- AAFinal_Sentence_Stimuli(Version 2).xlsx (Final sentence stimuli)
- Binary grammaticality instructions1.docx (Full text of instructions given to children completing the binary judgment task)
- Binary Judgement.zip (Zip file containing all video and audio stimuli, blank participant record and key sheets, and the sticker grid completed by children)
- Binary Judgement procedure.mp4 (Video illustrating the binary judgment procedure)
- Practice animations (Folder containing practice animations for the judgment warm up)
- Child instructions production.docx (Full text of instructions given to children completing the production task)
- CausativeAnimations.zip (Zip file containing all video and audio stimuli for the production task)
- JudgmentLists.zip (Zip file containing the different counterbalance lists for each language)
- Production procedure.mp4 (Video illustrating the elicited production procedure)
- Prereg Production and Binary Judgments.pdf (Preregistration of the methods used)

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).
References


Ambridge B: Argument Structure Overgeneralizations.docx. figshare. Figure. 2019. Publisher Full Text


Robenalt C, Goldberg AE: Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. Cogn Linguist. 2015; 26(3): 467–523. Publisher Full Text


Open Peer Review

Current Peer Review Status: 🛤️ rgba 🛤️

Version 1

Reviewer Report 14 May 2021

https://doi.org/10.21956/openreseurope.14078.r26696

© 2021 Pearl L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lisa Pearl
Language Science, University of California, Irvine, Irvine, CA, USA

I like to sign my reviews when possible. --Lisa S. Pearl

I think this manuscript makes great contributions in terms of addressing an interesting theoretical question about language acquisition: why children overgeneralize certain verb forms and how they recover from that overgeneralization. Moreover, the authors use multiple methods and look at data from multiple unrelated languages to assess the generalizability of their conclusions, and are pretty clear about the limitations of their current findings. As a computational cognitive modeler focusing on language acquisition, I especially appreciate the difficulties that go into (i) designing a computational model that connects concretely to empirical acquisition data, and (ii) interpreting model results in a cognitively-meaningful way. Because of this, I found myself deeply interested in what I would consider relevant modeling details, and unfortunately a bit confused by some of the main takeaways without those details. If possible, I'd love to see a revision that included some of that information, focusing on the aspects I mention in more detail below. Besides this, I had a few other specific thoughts that I discuss below.

(1) The model
(a) Model input, especially for children of different ages
It seems important to use input data distributions from children of the ages intended to be modeled (4;0-5;0, 5;6-6;6), as it's possible that input frequencies (particularly for uses of individual lexical items) would shift as children get older. Relatedly, children's perception of the relevant semantic properties may also be developing over time -- right now, adult data are used as a standin because that's the empirical data available (which is really great to have!). But that, along with the changing input frequencies, might explain why the model doesn't capture younger (immature) child behavior.

More generally, to capture children's knowledge at different ages with the kind of incremental model you have, I think you'd want to start the model off with some base level of knowledge (the equivalent of an informed Bayesian prior) that corresponds to what a child of that age is meant to know. For instance, if you wanted to pursue this idea, I could imagine running the model for the
4;0-5;0 child data, and then using that end state of that model as the start state for running the 5;6-6;6 child data (assuming the input frequency data actually differed between these two groups).

I don't think the current manuscript needs to do this, but I think it's worth discussing as a potential limitation and/or future work.

(b) Model implementation

(i) The manuscript notes that the model used is the very same one implemented by Ambridge et al. 2020, but it would be helpful just to give a very basic sketch of some of the finer details when it's first presented (e.g., how many input units are there, where the input data come from, etc.)

More generally, given the explanatory goals you have with the input features you're using, I'm curious about their motivation for using a neural network, rather than a more transparent classifier (like SVM, logistic regression) or cognitive model (like Bayesian inference). It seems like a more-transparent modeling methodology would speak to the goal of how predictive/explanatory the hypothesized features are. Of course, I realize you've already developed the model in the Ambridge et al. 2020 previous work, and want to test it here. But I do wonder if the explanatory goal would have been better-served by a different modeling choice. Perhaps the model incorporates information in a distributed way (e.g., like current word embedding approaches like GloVe or RoBERTa do), and these input representations would be hard to replicate in a different modeling type? At any rate, I do think a little more background on the modeling choice might be helpful somewhere in this paper for readers like me.

(ii) Split-half validation: I think it's good to see the split-half validation, even if you couldn't do it for K'iche' in Study 2. The results from split-half validation are much more believable, as opposed to the test-on-training when you use the full 60. That is, the split-validation captures what the model has learned more generally rather than what the model has learned (and potentially overfit) for these data. I know the full-60 correlations have higher values, but the split-half validation is more credible for the interesting explanatory claims you want to make about input frequency and the semantic features. Because of this, I'd be careful about playing up the full-60 modeling results compared with the split-half validation.

(iii) The discussion currently says about the modeling approach: “Given that an identical model can predict...without having been trained on any of these datasets”. I think I may have misunderstood something fundamental about the model then (and this may make my previous comments about the full-60 vs. split-half validation make less sense). I thought the model, as a neural net, gets trained to predict the correct output value on the basis of the input, and is given pairs of input-output to learn from over time. So, because the model has seen all 60 verbs (for the full-60 models) or 30 (for the split-half validation models), the model has in fact been trained (for the full-60) or partially trained (for the split-half) on these datasets.

If this isn't right, then I think the manuscript would definitely benefit from more description of the model itself and how it was trained, since this seems like a really important aspect of what you want to say in the general discussion. To me, the key idea is that certain aspects of the input have great explanatory power (with r around 0.5): the input frequencies of the forms, and the four semantic features. With these viewed as the relevant part of the input, a modeled learner can both overgeneralize and retreat from overgeneralization.
(c) What the model is meant to do: Related to the previous point, I think it may be helpful to draw out the explanatory power more of the frequency and semantic factors you investigate. If I'm understanding the goal of the modeling correctly, that's really what the model is meant to do: predict when overgeneralization does and doesn't happen, on the basis of these factors.

(2) What counts as overgeneralization
(i) When I was looking at Figure 3, I had a minor point of confusion about what counts as an overgeneralization. Overgeneralizations are defined as items where kids use the forms more equally than adults (i.e., kids think the forms are less different than adults). But, why not focus on the more natural overgeneralizations for the causative, where adults prefer the more-transparent (“A made B VERB”) over the less-transparent (“A VERBed B”) form moreso than children do?

In Hindi for example, it seems like overgeneralizations also include items where adults prefer the less-transparent (“A VERBed B”) over the more-transparent, because the adult difference score is positive.

(ii) Study 2:
The current manuscript suggests that less-transparent forms (“A VERBed B”) are more frequent in children's input, and that's why children use the less-transparent forms more than adults. Is the input difference in general for causative verbs, or just for individual verbs? That is, is the less-transparent more common for causative verbs as a whole, and that's what you think is causing the overgeneralization here? If so, then it means children have grouped together “causative verbs” as a class, and are tracking frequencies about that class.

If instead you mean that less-transparent forms are more common for individual verbs, then do you mean that "overgeneralization" (as defined here by using the less-transparent form more often than adults) is simply driven by the input? That is, it's just a reflection of an input that supports overgeneralizations, when defined as using the less-transparent form more often than you should. (Side note: This seems a little different from allowing a form that adults categorically think is not allowed. It might be helpful to note this.)

Followup if you meant children's input has more less-transparent uses for individual verbs: Is this something you could test for explicitly, by just seeing how well input frequency accounts for child judgments, and not including the 4 semantic features?

(3) Very minor thing: Figure resolution
Several figures in my pdf of the manuscript were rather fuzzy -- figures 1-3, an 9-11. It would be good to get better resolution versions of these.

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computational cognitive models of language development.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 04 May 2021

https://doi.org/10.21956/openreseurope.14078.r26710

© 2021 Kapatsinski V. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? Vsevolod Kapatsinski
Department of Linguistics, University of Oregon, Eugene, OR, USA

This article reports novel data on overgeneralization, a core topic in the acquisition of language. It expands a prior study by the researchers to a younger age range. An important strength of this paper is the cross-linguistic breadth of the investigation, which is unprecedented except for the paper’s companion piece (Ambridge et al., 2020). The main weakness of the paper is that the authors do not show that their computational model can account for the developmental trajectory. This is important because the main claim of the paper is that the model provides an account of *retreat* from overgeneralization. A second, related weakness of the analyses is that the authors say very little about how performance changes with age. Third, assuming that the model does account for the developmental trajectory, it would be important to show *why* it does. I elaborate on these issues below.

**Major points:**

1. **What properties of the computational model are important to account for the behavior?**

   The computational model used in this paper is a simple two-layer connectionist network, in which the input consists of a local (one-hot) encoding of verb identity, and 4 continuous semantic
parameters thought to be relevant to the choice between the causative constructions. Based on the preceding paper by the authors (Ambridge et al., 2020), the model also has an input node that represents whether the input is causative. The existence of this node is likely crucial for predicting that the more frequent causative construction will be overgeneralized, until the associations of specific semantic and lexical cues strengthen enough to override this initial bias. The output layer consists of three nodes for direct causative, indirect causative and 'other'. The learning rule is not described here, but is said to be a variant of Widrow-Hoff in Ambridge et al. (2020). It seems likely that the discriminative nature of this learning rule is crucial for the performance of the model, but this is not shown or discussed. It is also possible that a simple Hebbian learning rule would also do.

It is important to provide a full description of the model here so that the work could be replicated, and the paper could be read as a stand-alone piece. A full description should include: the learning rule, the activation function on the output node, learning rate, and any other parameters that were set. The authors also need to describe whether they have attempted to use different learning rules, activation functions, or parameter settings. Such explorations would be very informative for determining what properties of the model are responsible for its ability to explain the human behavior. In particular, does it matter that the learning rule is discriminative?

2. How does the behavior change with age?

It is not clear how the behavior in question changes with age. There are several possibilities, none of which are mutually exclusive. First, it could be that the children are more accepting of deviation from prior experience than adults (e.g., Kapatsinski et al., 2017). It could also be that children are gradually picking up on the semantic predictors conditioning the choice of the construction (Goldberg, 2019). Finally, it is also possible that, with age, children become more confident in their estimates of how individual verbs behave (e.g., suggested by Erker & Guy, 2012). Without knowing what changes with age, we cannot tell what the model should explain. I would like to see interactions between age and verb, and between age and the semantic predictors. According to p.20, "the main difference between 4;0-5;0 and 5;6-6;6 year olds is simply an across the board decrease in the production of overgeneralization errors, rather than any change in their by-verb patterning." I would like to see a statistical evaluation of this claim.

3. Does the model predict how the behavior changes with age?

Differences in the model's construction activations across verbs are shown to correlate with differences in ratings, judgments and production probabilities of both children and adults (in at least some languages). However, it is not clear how much of what the model is capturing here is variance shared between children and adults. That is, the model might be capturing semantic effects on construction choice that are equally robust in adults and children. If the model can account for retreat from overgeneralization, it is important to show that the model predicts how the behavior changes across age. The fact that the model does not show a better fit to kid data early in training and a better fit to adult data late in training is problematic if the correlations reliably change across development. If they don't, then the authors should show that the model captures what does change, even if this is only a simple increase in the use of the rarer construction with age.

4. How important are semantics, verbs, and the causative node?
Assuming the model can account for the changes in construction use with age, I would like to see what is responsible for those changes in the model. In particular, the model could be lesioned by removing the verb nodes, semantic nodes and/or the causative node, or injecting noise into the representations or the connections involving these different inputs. From the statement on p.5 that there is only a small decrement in performance when the model is tested on novel verbs, it appears that most of the model's performance comes from capturing the semantic influences on construction choice. Would removing the verbs altogether reduce the model's performance? The causative node seems necessary to predict overgeneralization of the frequent construction early on, though a bias node might also work for that purpose.

**Minor points:**

5. **Is there an interaction between directness and social desirability?**

It seems to me that the social desirability effect would not disappear from calculating difference scores (p.7). The two constructions (at least in English) differ in directness of causation, so it seems plausible that socially undesirable actions would favor the periphrastic construction (as in "caused to die" vs. "killed"). Does social desirability not correlate with constructional choice? Would it make sense to include it in the model as another cue?

6. **Are the same semantic dimensions relevant to all languages?**

There seems to be an assumption that the same four semantic dimensions should be relevant to the choice of the construction in all languages. I am curious as to whether this assumption is on a solid footing. It does not seem particularly surprising to me that there would be languages in which the choice of the causative construction is based on some variables other than the ones mentioned. Could this be the case in K'iche Mayan?

7. **More description of the training corpora**

More details on the training corpora are needed to evaluate whether they are representative of the input to children.

**References**


**Is the work clearly and accurately presented and does it engage with the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**
Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** My research deals with the role of domain-general learning mechanisms in language acquisition.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Zoe Ovans
1 Department of Hearing and Speech Sciences, University of Maryland, College Park, College Park, MD, USA
2 Neuroscience and Cognitive Science Program, University of Maryland, College Park, College Park, MD, USA

Yi Ting Huang
1 Department of Hearing and Speech Sciences, University of Maryland, College Park, College Park, MD, USA
2 Neuroscience and Cognitive Science Program, University of Maryland, College Park, College Park, MD, USA

Summary

In this work, the authors expand on a model presented in Ambridge et al., 2020, which predicts causative alternation acceptability on a verb-by-verb basis. This model takes as its input both corpus data as well as adults' semantic feature ratings for the verbs in question. In the present paper, the model is expanded to account for young children's overgeneralization errors (e.g. “I'm
dancing it” to mean “I'm causing it to dance”), both in offline comprehension and production. Notably, the authors attempt to account for children's overgeneralizations across five languages, and find that the Ambridge et al., 2020 model correlates with children's performance in most cases. These findings, they conclude, provide evidence that the discriminative learning model in question is a plausible explanation for how children retreat from verb argument structure overgeneralization errors, on mechanistic level.

Major Comments

To begin with, the 23 authors of this study should undoubtedly be commended for the large-scale collaborative effort this study represents! It's also exciting to see computational research that does not shy away from fine-grained cross-linguistic comparison. The question the authors target in this work, how do children retreat from overgeneralization errors, is timely and relevant to the field, and overall their methodology seems sound. That said, there are 4 major areas of concern the authors should address in order to both make their precise claims more intelligible in the current work and to make their conclusions more convincing:

1. Disconnect between the data presented and the research questions.

   a) The authors state in their abstract that “the present study demonstrates that a simple discriminative learning model ... constitutes a plausible mechanistic account of the retreat from overgeneralization.” (p2) and later that “the problem of how language learners come to appropriately constrain their argument structure generalization looks close to being solved“ (p22) because the model results correlate with both children and adults' causative judgment data and children's causative productions. However, it's not explicitly stated how the present model instantiates the mechanisms for learning argument structure. Looking at the 2020 paper, it seems that the answer is approximately that children are using a combination of lexical-semantic features, a causal/non-causal binary operator, and the ability to identify the particular lexical items to intuit which causative verbs ought to alternate which way. And this conclusion is reached because the model in question approximates human performance using these particular pieces of information as input. For this paper to function as a stand-alone work, these connections ought to be spelled out here as well.

   b) If the model presented is trained on the exact same data as the model in Ambridge et al., 2020, what changed between the two papers is just the values of the dependent measure the model is expected to approximate. If the model successfully predicts adult judgments, it may only explain child judgments to the extent that they mirror adult judgments. And these would be all the cases where children are not making overgeneralization errors. Therefore, to what extent does the model provide a mechanistic explanation for the retreat from overgeneralization? Perhaps a more direct measure of how well the model captures overgeneralization itself would be to correlate model performance, for each verb, with verb-level estimates of children's overgeneralizations (baselined to adults judgments).

   c) Furthermore, the 2020 paper did test the model on data from children, just slightly older ones (5-6 year-olds and 9-10 year-olds). The contribution of the current work is that the children tested were a bit younger still, at 4;0-5;0. The question this raises is: What are the data from these slightly younger children adding to our inferences about underlying generalizations? The authors state
that “the majority of [children's] overgeneralization errors are produced before [they're 5 years old]” (page 5). It would be helpful to more clearly spell out how testing children in the year before that allows us to make new inferences beyond those made in the 2020 paper.

2. Concerns about children's acceptability judgments and judgment data

a) The authors note that “the researchers who worked with the children reported that... children rated sentences, to some extent, on the basis of the social desirability of the events described” (p7). This, they reason, may have accounted for some undergeneralization errors such as children's low ratings for “someone broke the truck.” This raises questions about the validity of the grammaticality judgments and priming task in general. To what extent does the binary decision assessment conflate social desirability and grammatical acceptability? While the words in Figures 1-3 are a bit difficult to read, it seems possible that the more overgeneralized verbs correspond to more positively-valanced ones (e.g. dance, sing, play). It seems that the authors take the correlations with adult data to be evidence that children were attending to the grammatical acceptability on the whole, but it would be more convincing to show, say, a lack of correlation between child judgments and a valance measure.

b) The authors convincingly demonstrate that children's acceptability judgments correlate with those of adults. They take this to mean that “at least for English-, Hebrew-, Hindi- and K'iche'-speaking, children were indeed giving meaningful judgments” (p7). However, it should be noted that if children's responses are highly similar to adults', and the model used here has already been shown to reliably predict adults' judgement data, the mere correlation of model results and child judgment data should not be taken as a necessarily new finding, though this correlation appears to be the main result for Experiment 1. Simply put – if the model predictions correlate with adult data, and adult response data correlates with child response data, why wouldn't we expect the model predictions to correlate with the child data? Indeed, for the Japanese data where the child and adult responses do not correlate, the model does not appear to capture the child judgments. Additionally, the semantic feature judgments that serve as input to the model appear to be judgments from adults. It seems reasonable to assume that children's judgments will not match adults’ – even in this paper, there are instances where authors voice concerns about children not interpreting the valence of the sentences in an adult-like way (as just discussed). If the model is to be taken as a mechanistic explanation for how children retreat from their errors, it seems necessary to have its input parallel children's mental state as much as possible. The worry is that since the model relies on adult judgments, it might be ascribing to children knowledge that they do not yet posses. The model may therefore approximate children's binary judgment data for the wrong reason (it's relying on semantic information adults are privy to but children aren't). To avoid these concerns, the authors could provide some evidence for why they expect that children's and adults' semantic feature judgments would either match, or differ in a non-meaningful way.

3. Concerns about how to interpret the correlations.

The verb-level correlations from child and adult performance ignore the subject-level variability of estimates. If the data were analyzed instead with mixed-effects models (where verb is a random-effects variable), would adult ratings predict children's (and vice versa)
4. Concerns about the model data being run using adult corpus data

a) For both studies, it appears that the authors ran the Ambridge et al., 2020 model on causative alternation data gleaned from adult speech, and not speech to children. Given that the model is concluded to provide a potential mechanistic explanation for children's retreat from overgeneralization, it seems right to include causative alternation frequencies present in the speech that children themselves hear. Or at least, some evidence that the frequencies don't differ in the two types of speech is needed.

b) Finally, it's not clear from the present work that the semantic input nodes were necessary aspects of the model. Is it possible that the corpus frequencies just mapped on well to adults' and children's judgments? The opposite could also be asked – was the corpus frequency data needed in addition to the semantic features?

Minor Comments

1. Where possible, it'd be helpful to have vectorized images for the figures, or at least larger files. As is, it is difficult to read the individual words in Figures 1-3 and 9-10. Figure 11, while legible, is also a bit blurry. Once the figures are reproduced in higher definition it may additionally be useful to display the words in such a way that they don't overlap on Figures 1-3 & 9-10 (perhaps with a slight jitter).

2. Page 3-4 – A list of citations is given for papers and books that have investigated overgeneralization errors in English. While it's helpful to have a collection of citations for the past work on these errors, there is no explicit characterization of what this work has concluded. In order to show that it's relevant to the specific question of how learners overcome these errors, some curation of these citations is necessary.

3. Page 4 – A trio of theories are introduced here “preemption, conservatism via entrenchment (both statistical-learning theories) and verb semantics.” It seems to be a main goal of this work to show that these theories (or at least the unified version) provide the correct account of children's retreat from overgeneralization errors, so it would be helpful to give a precise definition for them here, as well as provide citations for their use. While this section is summarizing Ambridge et al., 2020, it becomes difficult to evaluate the main conclusions in the present paper on its own without (re)articulating these theories and how the model instantiates a combination of the three.

4. Page 11 – The authors make reference to the notion that “social desirability may be particularly salient in the more collectivist Japanese culture” in their explanation for the lack of correlation between the child and adult judgments for the Japanese data. While they provide a citation to this effect, this explanation appears to be a bit vague. In order to avoid being reductive in including cultural collectivism as a possible explanation, it'd be necessary to first establish that the cultures of other children in this study are less collectivist (and there is at least some evidence that this is a false assumption, e.g. Oyserman, Coon, & Kemmelmeier, 2002). Barring evidence to this effect, the cultural explanation doesn't convincingly explain the data pattern.
5. Page 20 – The authors state: “Comparison of Figure 9 (4;0-5;0) and Figure 10 (5;6-6;6) indicates that, by this later age, overgeneralization errors have all but ceased for English, Hindi and Japanese, and decreased considerably for Hebrew.” However, this difference is not particularly obvious from cross-figure comparison. A single figure comparing the results from both age groups on one graph (or at least closer together on a page) for each language would highlight this better.

Typesetting

For Figures 4-8 & 11, the labels could be cleaned up a bit (e.g. spaces added instead of underscores).

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My co-reviewer and I have expertise in language development and psycholinguistics, specifically in the area of argument structure.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.