PNAS

INAUGURAL ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Using computational modeling to validate the onset of productive determiner–noun combinations in English-learning children

Raquel G. Alhama[a,1], Ruthe Foushee[b,c,1], Dan Byrne[c], Allyson Ettinger[d], Afra Alishahi[e], and Susan Goldin-Meadow[c,f,2]

Affiliations are included on p. 9.

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2020.
Contributed by Susan Goldin-Meadow; received September 29, 2023; accepted October 7, 2024; reviewed by Julian M. Pine, Eva Portelance, and Charles Yang

Language is a productive system—we routinely produce well-formed utterances that we have never heard before. It is, however, difficult to assess when children first achieve linguistic productivity simply because we rarely know all the utterances a child has experienced. The onset of linguistic productivity has been at the heart of a long-standing theoretical question in language acquisition—do children come to language learning with abstract categories that they deploy from the earliest moments of acquisition? We address the problem of when linguistic productivity begins by marrying longitudinal behavioral observations and computational modeling to capitalize on the strengths of each. We used behavioral data to assess when a sample of 64 English-learning children began to productively combine determiners and nouns, a linguistic construction previously used to address this theoretical question. After the onset of productivity, the children produced determiner–noun combinations that were not attested in our sample of their linguistic input from caregivers. We used computational techniques to model the onsets and trajectories of determiner–noun combinations in these 64 children, as well as characteristics of their utterances in which the determiner was omitted. Because we knew exactly what input the model was trained on, we could, with confidence, know that the model had gone beyond its input. The parallels found between child and model in the timing and number of novel combinations suggest that the children too were creatively going beyond their input.

linguistic productivity | modeling language acquisition | grammatical development | generalization | syntactic categories

Language is a productive system. Although there are some utterances that are rotely learned and recalled as wholes (e.g., "How are you?"), there are many more productively generated utterances that are not likely to be produced a second time. Take the following example, which first appeared in the Letters to the Editor of a popular TV magazine [quoted in (1)]: "How Ann Salisbury can claim that Pam Dawber's anger at not receiving her fair share of acclaim for Mork and Mindy's success derives from a fragile ego escapes me". This convoluted utterance is completely grammatical, and it is also understandable—because we know how English works, not because we have memorized the utterance. Here, we tackle the onset of productivity in children learning determiner–noun combinations in English and explore an approach that can tell us when a learner produces pattern-conforming utterances.

The question of linguistic productivity has been at the heart of a theoretical debate in language acquisition—do children come to language-acquisition equipped with abstract categories that they deploy at the earliest moments of language learning, or do they construct these categories (2–8)? Our goal is not to resolve this debate, but to provide a unified behavioral/computational approach to the onset-of-productivity question, which is central to the debate—how can we determine when a child has achieved linguistic productivity?

One approach to knowing when a pattern is productive in a language-learning child is to look at the errors the child makes. For example, at a certain point in development, children learning the endings of English verbs will occasionally say "I eated that yesterday." Although this is not an acceptable utterance in English, the utterance does reveal that the child who produced it knows an English pattern—that *-ed* is added to verbs to indicate the past. Errors that reveal knowledge of a pattern can only occur when there are exceptions to the pattern; in this case, *ate*, the correct form, lacks the *-ed* ending and thus is an exception to how past tense is regularly formed in English. These exceptions give learners an opportunity to create a form that they have not heard before but still follows an English pattern. The problem

Significance

A difficult problem in describing language acquisition is knowing when children go beyond their input to produce novel, structured utterances—that is, to achieve linguistic productivity, the hallmark of human language. We address this problem by detailing onsets and trajectories of 64 English-learning children producing determiner–noun combinations (*the dog, a dog*) and by capturing these behaviors with a computational model. Because we know the model's input, we can determine when it predicts combinations not in its training set. We find parallels between child and model in the timing of novel combinations, suggesting productivity in the children. Marrying behavioral observations and computational modeling provides an approach that can be used to assess productivity in any language, spoken or signed.

Author contributions: R.G.A., R.F., D.B., A.E., A.A., and S.G.-M. designed research; R.G.A., R.F., and D.B. performed research; R.G.A., R.F., and D.B. contributed new reagents/analytic tools; R.G.A., R.F., and D.B. analyzed data; and S.G.-M. wrote the paper.

Reviewers: J.M.P., University of Liverpool; E.P., HEC Montreal; and C.Y., University of Pennsylvania.

The authors declare no competing interest.

[1] R.G.A. and R.F. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: sgsg@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2316527121/-/DCSupplemental.

Published November 21, 2024.

with this approach is that it is applicable only in forms that violate regular patterns of a language.

But the approach is based on a useful assumption—if learners produce pattern-conforming forms without ever having heard them, they must have an understanding of the patterns that generate the forms. Language acquisition researchers routinely use this reasoning when they uncover apparently productive uses of a form in a child and can find no evidence that the child has heard this form before. If, for example, a child hears *a pineapple* and then produces *the pineapple* without ever having heard the combination, we then guess that this combination was the child's creation. But, of course, our sample of the child's input is just that—a sample. Can we be certain that a child has never heard a *particular* well-formed combination?

Here, we add computational modeling to our behavioral analyses to address this problem. We use a model that simulates the onset and developmental trajectory of a productive form in a large sample of English-learning children. We know all the utterances that the model received during its training. As a result, we can tell with certainty whether, and when, forms the model predicts go beyond the input given. If the generalized forms predicted by the model are comparable to the generalized forms produced by children, we greatly strengthen the evidence that these forms are not rotely learned in the child and instead are productive forms.

We use determiner–noun combinations in English (*a dog*, *the dog*) as a test case for our approach for two reasons. First, this construction has been a focus of research designed to address the abstractness of linguistic categories (2–8). Second, the determiner–noun construction is regular[*] in English and thus contains no violations of the pattern, forcing us to seek new types of evidence for productivity. Children begin to use *a* and *the* with nouns early in development. Our question is when do English-learning children know that a noun can be used with both *a* and *the*; that is, when do they have productive use of determiner–noun combinations?

Background on Behavioral Measures and Computational Models of Determiner–Noun Productivity

In seeking a quantitative measure of determiner–noun productivity that could be applied to spontaneous child productions, Pine and Lieven (2) developed an overlap score. The overlap in question is between the set of nouns used with *a* and the set of nouns used with *the*. For example, a child who produced *a dog, a sock, the dog,* and *the plant* would receive an overlap score of 1/3, or 33%, having produced one noun type out of three (*dog, sock, plant*) with both *a* and *the* (*a dog, the dog*). But using a measure based on all the nouns a child produces misses an important aspect of English use—certain nouns in English are more likely to appear with one of the two determiners (e.g., *the bathroom* is more frequent than *a bathroom*). As a result, some nouns are not likely to occur with both determiners in spontaneous discourse. Yang (8) solved this problem by computing an *expected* overlap score—the overlap score that children would receive if they used language like adults (i.e., a score that takes into account how likely a noun in adult speech is to occur with both determiners). When applied to a sufficiently large corpus, children's *expected* and *observed* overlap scores ought not differ *if* they have a productive use of determiner–noun combinations.

Meylan et al. (9) took a different approach to the problem and simulated determiner–noun productivity in a Bayesian model. Under their model, a child's determiner productions for each of their nouns are guided by two information sources—(A) direct experience and (B) productive knowledge. The strength of each source's contribution to the child's productions is determined by individual weighting parameters. Meylan et al. applied their model to data from 27 children, the youngest of whom was 9 mo. Their simulation results suggest that the youngest age groups in previous behavioral studies were not young enough to reveal a gradual increase in productivity. In addition, they argue that because previously used measures are highly sensitive to the size of the sampled data, the measures ought not be used to estimate productivity at the earliest stages of learning when children do not produce many determiner–noun combinations.

Here, we follow Cartmill et al. (10), who used a straightforward measure that works well with small samples to identify when children first produce a particular construction. Onset age for productive determiner–noun combinations in our study is established when a child uses both *a* and *the* with at least two different nouns (*a dog, the dog, a book, the book*). We first establish the onset age for determiner–noun productivity in 64 English learners. We then use a computational model to simulate onset age and trajectory of determiner–noun combinations in each of these children.

Previous computational studies have simulated determiner productivity using neural network models. Phillips and Hodas (11) used an autoencoder architecture, whose goal is to reconstruct (or repeat) an input utterance. The architecture was trained on child-directed utterances from corpora in CHILDES (12). The model learns a compact, latent representation for every incoming utterance, which it then uses to regenerate the same utterance. The authors measured the estimated and empirical overlap scores in adult utterances from the training corpora and in the utterances generated by the autoencoder model. They then showed that if the model's parameters are set to allow for more generalizability, its estimated overlap scores are close to those of adults. Crucially, this study does not compare the behavior of the trained model to the behavior of children and therefore says little about the trajectory children follow in learning determiner–noun combinations.

To address both issues, Alhama et al. (13) compared the predictions of a neural network model to longitudinal observations of children producing the determiner–noun construction. The model, which is based on the Transformer architecture (14), is trained to predict masked words in a sentence. They found that the model mimicked *overlap* scores and the simpler *onset* metric (10) in both the Manchester corpus (4) and the Language Development Project corpus [LDP (15)] used here (see next section), following developmental trajectories comparable to the children's in both corpora. Here, we take a step forward by using the same model metric to further analyze when child and model go beyond the data given.

The contributions of our behavioral and computational study are three-fold. i) We characterize the trajectory of determiner–noun combinations in individual children and, in so doing, confirm and extend to a larger group of children Meylan et al.'s (9) finding that determiner–noun productivity emerges gradually and stabilizes with age. ii) We show that the computational model developed by Alhama et al. (13) not only closely mimics children's productive uses of determiners over time, but also captures the uncertainty surrounding determiner omission in their early utterances. iii) We take advantage of our fully comprehensive knowledge of the model's input data (something that is not feasible in children) and identify cases in which the model makes predictions that yield *novel* determiner–noun combinations, not found in its training set. The children show comparable novel combinations when their data are compared to their parents' data. We then use

[*] Mass nouns (e.g., *flour*) complicate the determiner–noun pattern in English; however, previous work on this topic has omitted this exception for simplicity.

the onset of novel determiner–noun combinations to validate our measure of determiner–noun productivity (when *a* and *the* are both used with at least two different nouns) in the child and in the model. In the child, the age at which children produce their first novel determiner–noun combination correlates with the age at which they meet our criterion for determiner–noun productivity. In the model, the session when the model predicts novel determiner–noun combinations (i.e., a combination not found in its training set) correlates with the session when the model predicts *a* and *the* in the context of at least two different nouns, thereby providing evidence that this metric captures productivity.

Observing and Modeling Determiner–Noun Productivity in Children Learning English

Our Behavioral Data. The behavioral data for this study come from the LDP corpus (LDP, see ref. 15) in which 64 English-learning children were observed longitudinally. Children and their primary caregivers were video-recorded while engaging in spontaneous interactions for 90 min in their homes every 4 mo, from 14 to 58 mo. We used spontaneous production data from each of the children and their caregivers (the parent in every case). The parent data served two purposes. First, we compared each parent's data on determiner–noun combinations to the data from their child. Second, we trained our model on child-directed speech from the LDP corpus (see next section).

To estimate determiner–noun productivity, we follow Cartmill et al. (10) in requiring that the child produce two instances of the relevant construction. Cartmill et al's goal was to compare the onset of point+noun combinations (e.g., point at dog + *dog*) to the onset of determiner–noun combinations (e.g., *the dog*). Their criterion for the onset of determiner–noun combinations was met when a child produced *a* and *the*, each combined with two different nouns (e.g., *a girl, a bottle, the dog, the cookie*; the authors were not interested in our productivity question—whether the child knew that any noun used with *the* can also be used with *a*, and vice versa). They found that the age at which a child first produced these point+noun combinations reliably predicted the age at which the child first produced determiner–noun combinations. Moreover, the point+noun combinations declined in frequency after the onset of determiner–noun combinations, validating their onset criterion. Here, we adopt this two-instance criterion and assume that a speaker (child or parent) demonstrates *productive* use of determiner–noun combinations when the speaker uses both *a* and *the* with the same noun, and does so with at least two different nouns (*a car, the car, a bottle, the bottle*).

Our Computational Model. Our computational goal is to use a model, trained on child-directed data, that simulates the developmental trajectory of determiner–noun productivity. We do not seek to capture the mechanism that children use for learning and processing language. Nevertheless, we chose a modeling framework and architecture that satisfied two criteria. First, the model must not rely on any data or supervision signal that is not available to children. Second, the model must not rely on any explicit, latent representation of abstract syntactic categories in advance. However, our model does have an advantage over child learners—the model can use the training data over and over; in contrast, the child's language-learning process is incremental and online (although children can, in principle, revisit what they have heard and thus process it multiple times).

A common goal, inspired by human language processing and widely used in building cognitive computational models of language, is for a model to predict the next word in an incoming utterance (known as *language modeling* in computational linguistics). There is strong empirical evidence that both children and adults form expectations about incoming words when receiving and processing an unfolding utterance. This task is simulated in various (mainly neural network-based) modeling architectures, going back to Elman (16, 17). Our focus here is on production rather than reception. As a result, we use a model that makes predictions about determiners based on the context of the child's entire utterance since children have access to this context when they produce their utterances. The Transformer-based model used by Alhama et al. (13) meets our two criteria. The model is trained on linguistic data that the parents in the LDP corpus produced; hence it does not receive input that is different from a child learner's input, meeting our first criterion. In addition, the model does not rely on preexisting syntactic knowledge, which meets our second criterion.

We trained the model from scratch on child-directed utterances from the LDP input corpus. Since the child-directed data for each individual child were not large enough to train the model, we accumulated utterances from all the parents in LDP, divided by the observation session. We incrementally trained the model on these data so that at each stage in learning, the model saw all the child-directed data up to that point. The model was trained to predict the masked words in an utterance using an error-driven algorithm.

The model was tested on individual child-produced utterances taken from each observation session. We first extracted all the determiner+noun usages in the utterances produced by each individual child, following Pine et al. (3, 18). We then masked the determiner in each child's usage and fed the usages to the model to predict the most likely filler for the masked slot. As an example, for the utterance *Where is the stroller* in the child data, we present the model with *Where is [MASK] stroller*. For each masked slot, we replace the child's word with the word predicted by our model. We use the resulting utterances to determine when the model first predicts two different nouns, each produced with both *a* and *the*, for each child.

Study 1: Identifying the Onset of Determiner–Noun Productivity in Children and the Model

The Children. Children first produced a determiner, either *a* or *the,* between 14 mo (the first observation session) and 38 mo; mean onset age = 21.80 (SD = 4.83) mo. Of the 64 children, 63 met the criterion for determiner–noun productivity within the 12 observation sessions; that is, they produced *a* and *the* with the same noun and did so for two different nouns. The number of sessions between a child's first determiner and achieving determiner–noun productivity varied from 0 sessions (n = 4) to 7 sessions (n = 1); mean number of sessions between first appearance of determiners and determiner–noun productivity = 2.29 (SD = 1.36). In other words, after producing their first determiner, children took, on average, 9 mo to become productive at approximately 30 mo (recall that the interval between each session was 4 mo; *SI Appendix*, Table S1, for descriptive data on the corpus).

Did children have the opportunity to meet our productivity criterion before achieving it? To find out, we examined the sessions prior to the session at which each child first met our criterion, and asked whether the child had produced at least two different nouns twice (e.g., *shoes* twice and *book* twice) during this period. We found that 53 children (84%) had produced two nouns two times in at least one session preceding that child's onset of productivity. In other words, the children had produced enough noun phrases that they *could* have met our onset productivity criterion—but they did not, suggesting that we had captured the period when children were first productive. For the remaining 10 children (16%), their first productive session was the first time that the child met the enabling conditions for productivity.

Fig. 1 presents the median number of different nouns (i.e., noun types) that appear with both *a* and *the* determiners (orange dots), grouped according to the age at which the child first met our criterion for productivity. The dashed line represents the lower boundary of our criterion for productivity (i.e., two nouns, each produced with *a* and *the*). Note that after having met the criterion, children in each group produced roughly the same number of different nouns with both *a* and *the,* no matter when they first achieved productivity. The majority of children (52 of 63, 83%) met our productivity criterion on at least half of the sessions following their onset; 22 children reached this criterion on every subsequent session. The emergence of determiner–noun productivity in children is a gradual process and, for most children, once the generalization criterion is achieved, it is maintained.[†]

However, the onset measure is sensitive to sample size, which generally increases as children grow older and become more talkative. Thus, meeting our onset criterion (*a* and *the* both used with at least two different nouns) might merely reflect the fact that the number of determiners children produce increases over developmental time. To address this concern, we looked at the sessions prior to the one when the child reached our criterion (the onset session) and took the maximum number of determiners produced by the child during one preonset session. The maximum number ranged from 4 to 287 (M = 59) across children, which itself suggests that onset of determiner–noun productivity is *not* due to the number of determiners produced (16 children had to be eliminated from this analysis because they did not produce the minimum of 4 determiners in their session prior to onset).

We then used the maximum number of determiners produced preonset and randomly selected this number of determiner–noun combinations from those produced by the child at the onset session. We then calculated whether the selected combinations met the criterion for productivity (two nouns each used with *a* and *the*) in the truncated sample; if so, the session was considered productive. We conducted 100 simulations for each child's onset session and calculated how likely the child was to reach our productivity criterion for this session. We found that when the number of determiners produced at onset was limited to the number produced prior to onset, the probability of meeting our productivity criterion was 0.20 (SD = 0.29), compared to the preonset level, which was 0. In other words, the children display productivity at their onset session even when we restrict their number of determiners to preonset levels.

The Model. Fig. 1 also presents the number of noun types that appear with both determiners in the model's output for each session (blue lines). The model's predictions for each child create a pattern of onset and maintenance of determiner–noun productivity comparable to the child data.

Fig. 2, *Left* panel, highlights the parallels between model and child by plotting the median number of noun types occurring with both determiners (*a* and *the*) produced by all children (orange dots), predicted by the model (blue line), and produced by the children's parents (magenta dots). Not surprisingly, the parents were fully productive from the first recorded session. On average, both children and the model meet our criterion for productive determiner–noun combinations at 30 mo (±4 mo).

A second way to look at the fit between model and child is to correlate onset sessions for the two. The age at which children first met the onset criterion strongly correlates with the onset session determined by model's predicted responses, r = 0.71, as shown in Fig. 2, *Right* panel.

However, it is possible that the similarity between the model's predictions and the children's behavior is a function of properties of the child-produced utterances that we used at test (rather than true similarities between determiner–noun patterns for child and model). To address this concern, we ran a control experiment in which we tested the model on parent-produced utterances. We sampled utterances with determiner constructions (n = 48 for each session) separately for the child utterances and for the parent utterances; we then used them as test frames for the model. We computed the number of different nouns produced with both *a* and *the* for each set of sentences and performed a paired statistical test. We repeated this experiment 10 times; none of the tests yielded a significant difference between test frames taken from parent speech and test frames taken from child speech. This result rules out the hypothesis that the parallels between model and child are due to the linguistic context of child produced speech used at test. Note that the analysis also controls for sample size (i.e., number of test items given to the model for each session).

To further address concerns about the impact of sample size on our onset criterion, we systematically undersampled determiner–noun combinations in the model just as we did for the children. As in the behavioral analysis, we took the number of determiners that each child produced prior to the session when that child achieved productivity, and randomly selected that number of determiner–noun responses in each child's onset session to use as test items for the model. We asked whether the model's predictions passed our productivity criterion in this randomly selected sample, and conducted this analysis 100 times per child. We found that the model predicted responses that displayed productivity on 0.18 of the randomly generated samples, which is comparable to the 0.20 found for the behavioral analysis, and, even more important, is above 0. Thus, the *model* can make predictions that suggest determiner productivity even when we limit the number of items it is tested on to the number the child produced prior to onset when productivity was 0.

Note that we are not suggesting that our model is capturing the processing steps that children follow to arrive at determiner–noun productivity. Rather, we are using the model as a tool to explore the trajectories individual children follow in acquiring determiner–noun constructions. It is likely that many types of models can, in principle, serve as an assessment tool for child productivity since many data-driven models can generalize and produce novel strings. Importantly, however, not all models account for our behavioral data. We used an n-gram model with backoff (which has the advantage of being online and incremental, as in child language learning) to predict our behavioral data. Nevertheless, we found that compared to the children and to the model that we used in our study, the n-gram model underestimates productivity (*SI Appendix*, Fig. S1).

[†] Although children tend to remain productive after onset, their productivity levels do dip on occasion, as do the parents' levels. To get a handle on these fluctuations, we calculated type/token ratio (number of different nouns/total number of nouns) for nouns the child produced in each session. We correlated this ratio with the number of productive nouns the child produced in the session. We did the same for each parent's nouns to explore fluctuations in their data (see Fig. 2, *Left* graph, magenta dots). In both groups, type/token ratio increased as the number of productive nouns decreased (correlation in children = −0.33 post onset; correlation in parents = −0.51). We then looked at the sessions where the children's (and the parents') number of productive nouns decreased and calculated the proportion of those sessions that were accompanied by an increase in type/token noun ratio. We found that 119/178 (67%) of sessions where there was a decrease in the number of productive nouns were accompanied by an increase in type/token ratio for the children; 186/289 (64%) for the parents.

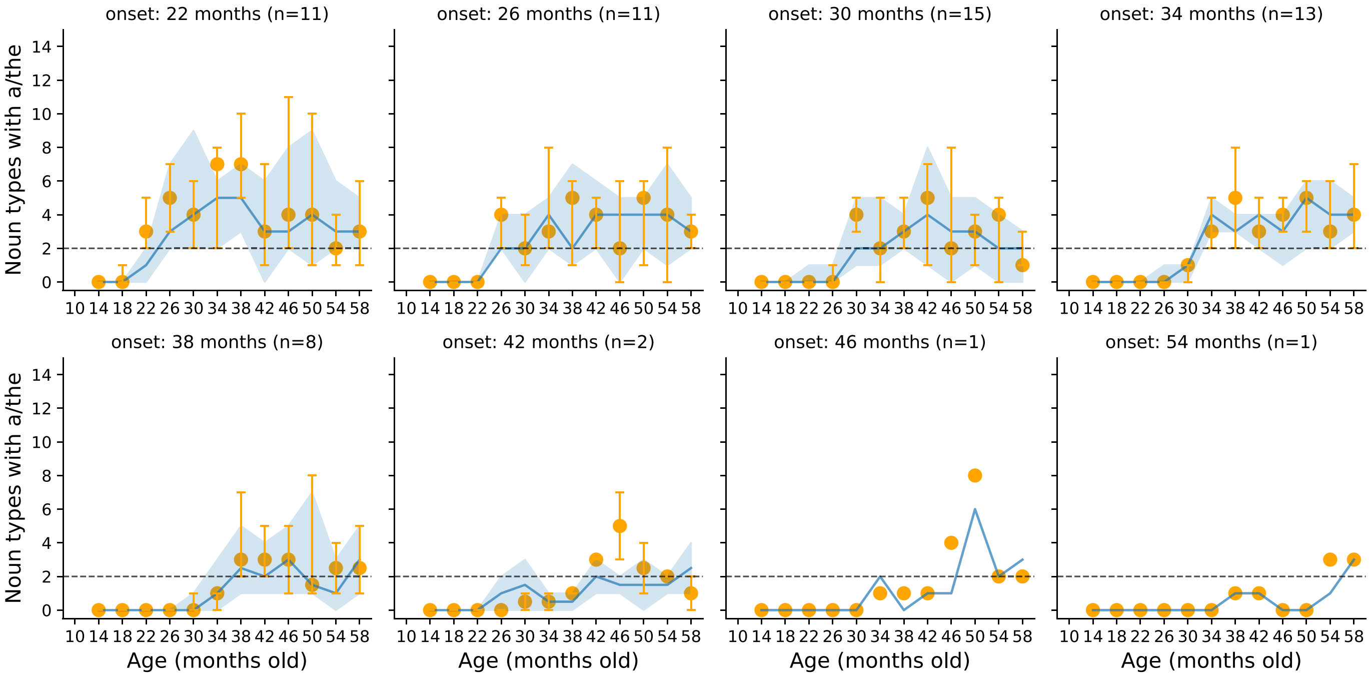

Fig. 1. Median number of noun types produced with both determiners, grouped by the age at which the child (orange dots) first achieved productivity or the model (blue line) first predicted two different nouns combined with *a* and *the*. The dashed horizontal line denotes two different nouns, each used with both *a* and *the*.

Study 2: Modeling Uncertainty Surrounding Determiner *Omissions* in Children's Utterances

Our model does a good job of predicting determiners in utterances where children produce them, increasing in accuracy with child observation session (*SI Appendix*, Fig. S2). But the model was not given utterances where children should have produced a determiner but did not. The model might therefore predict determiners when children omit them. Not capturing children's omissions, particularly at the early stages, weakens the parallels between child and model.

To address this issue, we enlarged the set of utterances used to test determiner productivity in the model. In addition to the utterances children produced that contained a determiner, we added children's utterances that should have contained a determiner but did not (e.g., "open bottle," "cat eating"; see experimental setup in *Materials and Methods*). We masked the position where the determiner should have been, and gave the model these determiner-omitted utterances, along with the children's determiner-produced utterances. The model does not have the option of leaving a blank so (unlike the child) it must fill in a word for the masked item. If, however, the model is working from a system that is comparable to the child's, it should be relatively uncertain about the word it produces in the blank in the test utterances that come from the children's nonproductive period; and it should become more certain as the children's systems become productive. We estimate uncertainty by measuring the model's entropy, with higher entropy reflecting more uncertainty.

In practice, the model predicts a probability distribution for the lexical items that could fill in the slot of the missing word. In Study 1, we focused on the most likely item from this distribution and showed that it paralleled the children's use of determiners. In Study 2, we compute the entropy over the predicted probability distribution. If the probability mass is more peaked and centered on a few lexical items, then the model has lower entropy and less uncertainty; if the probability mass is more evenly distributed,

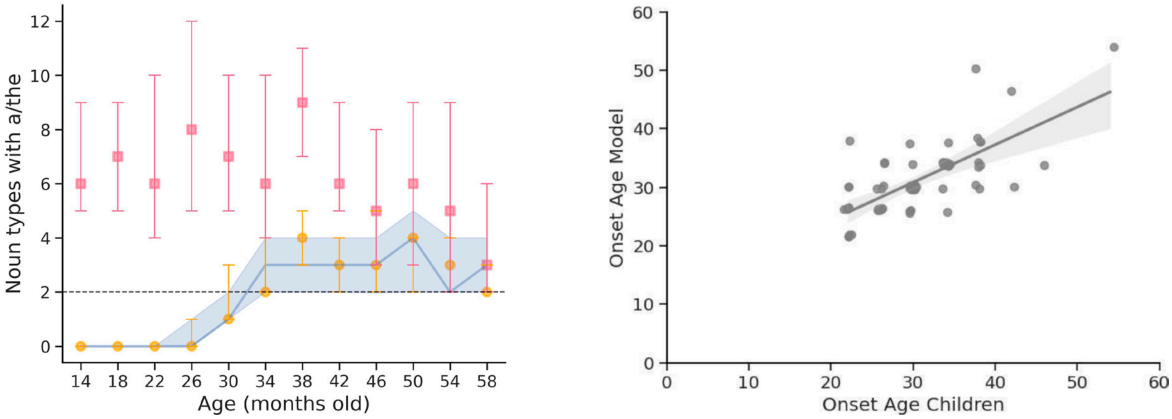

Fig. 2. *Left* graph: Median number of noun types combined with both *a* and *the* that the children produced (orange dots), the parents produced (magenta dots), and the model predicted (blue line). The dashed horizontal line denotes two different nouns, each used with both *a* and *the*. *Right* graph: Age at which children first produced two different nouns, each with *a* and *the* (*x*-axis) and the session at which the model predicted two different nouns, each combined with *a* and *the* (*y*-axis). Pearson's correlation is r = 0.71. A random jitter of 0.5 has been added to overlapping points in the graph. The *Right* graph is reprinted from (13).

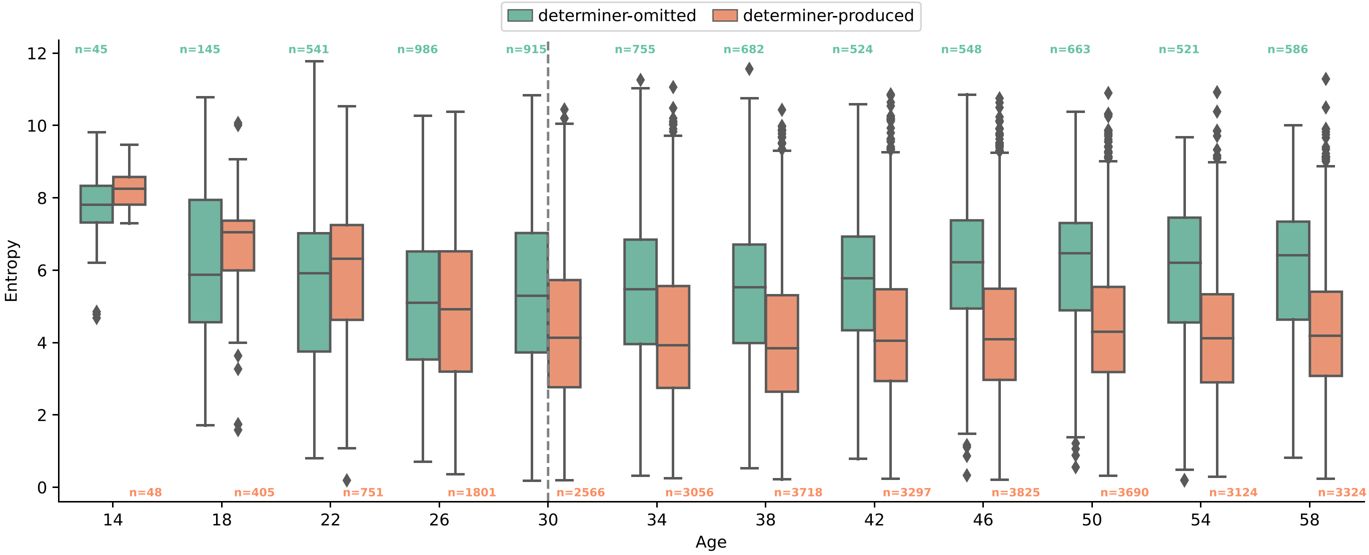

Fig. 3. The mean entropy score for the model at each developmental session for utterances in which children correctly produced a determiner (determiner-produced) or failed to produce a determiner when it was needed (determiner-omitted). The boxes show the quartiles; the horizontal line in the box is the median, and the vertical line extends to the rest of the distribution. Diamonds indicate outliers. The vertical discontinuous gray line indicates the average age at which children in Study 1 began to productively use determiner–noun combinations; note that the line coincides with the model's stabilization of entropy.

then the model has higher entropy and more uncertainty. We expect the model to be relatively uncertain (higher entropy) on the items from the children's early nonproductive sessions, for both the determiner-omitted and determiner-produced utterances. As the children's systems become productive, we expect uncertainty to decline, again for both types of utterances.

Fig. 3 presents the model's mean entropy scores for determiner-omitted utterances and for determiner-produced utterances. Note that the entropy score for both types of utterances begins high and, as expected, decreases and stabilizes around 30 mo, when the children became productive users of determiner–noun combinations. Interestingly, after this point, the mean entropy score for determiner-omitted utterances is consistently higher than for determiner-produced utterances, perhaps because the places where children fail to produce determiners later in development are generally less obvious.

Study 3: Going Beyond the Input Given to Children and the Model

We are assuming that the children are productively generating their determiner–noun combinations. But they could be copying these combinations from their parents, particularly since we require combinatoriality with only two different nouns to achieve onset. Our next step is to examine child combinations in relation to the input children receive. We use the child's first instance of a novel determiner–noun combination to validate the child's productivity onset measure; we do a comparable analysis for the model.

Fig. 4 displays the number of novel determiner–noun combinations produced by the children (in orange) and predicted by the model (in blue). A combination was considered novel in a child if the combination did not appear in that child's parent's input up to that point. A combination was considered novel in the model if it did not appear in the input that the model had been trained on up to that point; recall that the model had to go beyond the input from *all* the parents to be credited with a novel combination. The figure shows the age at which the child produces and the model predicts its first novel determiner–noun combination. We may have overestimated the number of novel combinations in the children simply because we have only a sample of their input; they could easily have heard, at other points in their lives, some of the combinations we labeled as novel. However, since we have access to *all* the training data the model has received in its lifetime, we can track novel combinations with confidence.

Our last step is to use the onset of novel determiner–noun combinations to validate the productivity onset measure for the child. If the first novel instances of determiner–noun combinations coincide with the onset age identified using our productivity criterion, we can take this as evidence that the onset measure is a valid measure of productivity. Fig. 5, *Left* graph, shows the correlation (r = 0.64) between age of productivity as estimated by the onset measure (*x*-axis) and age of first *novel* determiner–noun combination (*y*-axis) for the children. The positive correlation validates our onset measure as an index of linguistic productivity for the child. Fig. 5, *Right* graph, shows a parallel correlation (r = 0.58) for the model. The positive correlation not only validates the onset measure for the model, but does so with a novelty measure that is more reliable than the child's (since we know precisely which combinations the model was trained on). The parallel findings lend weight to the claim that the child, like the model, has gone beyond the data given.

Discussion

We have taken an approach that marries computational modeling and behavioral analysis to identify the onset of linguistic productivity in child language learners, using determiner–noun combinations as a test case. We used a large, longitudinal corpus of spontaneous speech to examine the onset of productive use of determiner–noun combinations in children learning English. We used a straightforward measure of productivity (to be productive, a child had to use *a* and *the* with the same noun, and do this for two different nouns), and were able to identify the age at which 63 of 64 children began to productively produce determiner–noun combinations. We found that on average, children began to produce productive determiner–noun combinations at approximately 30 mo, roughly 9 mo after they produced their first determiner, confirming previous findings (cf. 9). This delay is particularly striking given comprehension studies showing that 14-mo-old Canadian French-learning children (19) and 14- to 16-mo-old German-learning children (20) can use the determiners they hear to categorize a following novel word as a noun,

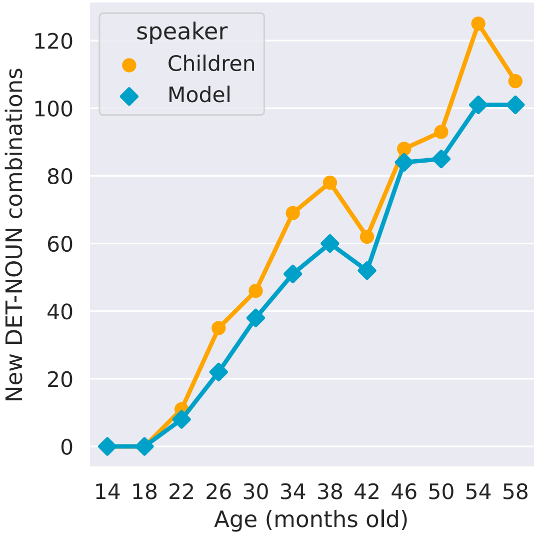

Fig. 4. Number of determiner–noun combinations produced by the child (in orange) and predicted by the model (in blue) that were not in the child's parental input or the model's training set up to that session, and that also had not been earlier produced by child or predicted by the model.

and they are able to do this many months before producing their first determiner. The open question that our findings highlight is why, given this early understanding that determiners precede nouns in *comprehension*, children fail to demonstrate productivity when they first *produce* determiner–noun combinations and take, on average, 9 mo to do so.

We then made use of the computational model we developed that was trained on parent input data to all 64 children (13). The model's learning trajectory of determiner–noun combinations paralleled the children's onset age and trajectory in the determiners that children produced, and displayed uncertainty at the appropriate sessions in the determiners that children failed to produce. Importantly, the model that we used was not pretrained on determiner–noun categories, but instead was trained on the linguistic input provided to the 64 children in our sample. Our findings thus suggest that these abstract categories *can* be derived from linguistic input. We return to this point later in the discussion.

Our final step was to leverage the parallels that we found between child and model to explore whether children truly go beyond the input they are given. The children in our study did indeed produce determiner–noun combinations that could not be found in the linguistic input we had in our sample. But, of course, these parental utterances are only a small sample of the language children actually hear. We also found that our model predicted determiner–noun combinations that could not be found in the data on which it was trained. In this case, however, the training data are all the input the model received. We can therefore be certain that our model has gone beyond the input given, predicting previously unseen combinations that meet our criterion for productivity. The parallels that we see between child and model lend weight to our claim that the children have gone beyond the input given to achieve linguistic productivity.

Our model allows us to ask whether a preexisting syntactic category is needed to develop the determiner class and use it productively with the noun class. Our results suggest that the determiner category *can* be learned in a bottom–up fashion, implying that a priori abstract categories are not a prerequisite for explaining determiner–noun productivity. There is, however, an important caveat. Our model had access to the child's entire utterance (not just the noun). This decision is not unwarranted since speakers (including children) know which words they are about to produce. But a consequence of this decision is that words other than the noun may have influenced the model's prediction. For example, repeated occurrences of *give me the <noun>* may prompt the model to learn that *the* can be used after *give me*. Note, however, that the same point can be made for children. As a result, the notion of productivity that we have modeled is determiner production that may be cued not only by the noun that follows it, but also by words, such as the verb, that precede it. However, since our goal is to determine when the model predicts that both *a* and *the* will appear in the slot preceding the *same* noun, the noun must be centrally involved in the prediction.

The naturalistic input from the children's parents that we used to train the model was sufficient to lead to model predictions that reflect determiner–noun productivity. But we do not know how little input, and what type of input, are necessary for the model to predict these productive forms. Our model would obviously not be able to predict determiner–noun combinations if it did not have access to the linguistic input we provided, but children *are* able to do so. The evidence comes from *homesigners*—profoundly deaf children whose hearing losses prevent them from acquiring a spoken language, and whose hearing parents have not exposed them to a signed language. Despite their lack of linguistic input, these children communicate using self-generated homesigns. Although homesigns do not display all the properties of natural language (21, 22), they do contain productive (23) determiner–noun combinations (24). Importantly, the homesigners' hearing parents do *not* display determiner–noun combinations in the spontaneous gestures that they produce when they talk to their deaf children (25), indicating that the children do not have a model for this category either from a conventional language or from spontaneous gesture.

The homesign observations suggest that children come to language learning prepared to create a determiner category if they are not receiving linguistic input. Our behavioral findings here make it clear that if children *do* receive linguistic input, they can use that input to achieve productivity in determiner–noun combinations early in development, although it does take them approximately 9 mo after they have produced their first determiner to achieve determiner productivity in the language they are learning.

How little input can we give our model and still have it achieve determiner–noun productivity (e.g., how few nouns need to be paired with both *a* and *the* for the model to predict utterances consistent with productivity)? Our methodology provides us with a starting point for further investigating the amount and type of input that triggers learning the determiner category. Our first step was to characterize at a fine-grained level the timing associated with onset and development of determiner–noun productivity in children *and* in a computational model. Having completed this step, we can now ask questions about the type and amount of linguistic input needed to model the individual learning trajectories displayed by the children in our study. For example, our analysis of child-produced determiner–noun combinations in a sample of 64 children indicates variability in the onset and rate of productivity among the children. We can use our computational model to investigate factors that might have led to these differences (e.g., characteristics of the linguistic input each child receives during their learning process, the patterns of interaction between parent and child, etc.). In the current study, we combined child-directed data for all the children in our corpus to train our computational model (we individualized the data by using child-produced utterances as test cases for each version of the model). We made this decision because we did not have enough training data for each child. In the future, we hope to develop creative ways to simulate differences in the linguistic input that individual children receive (including

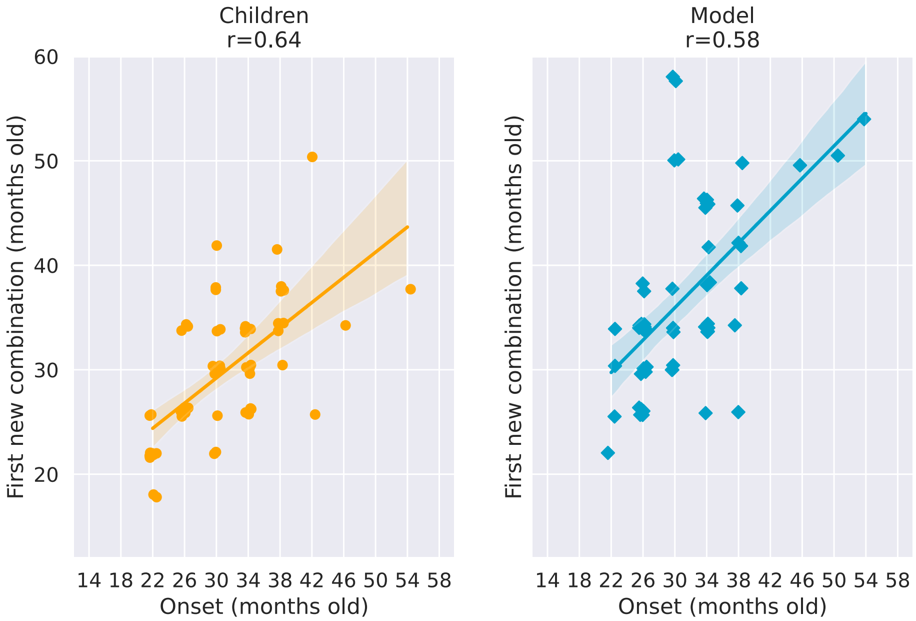

Fig. 5. *Left* graph: Correlation between the child's age at onset of determiner–noun productivity and the first novel determiner–noun combination not in that child's parental input up to that session. *Right* graph: Correlation between the session when the model first predicted two different nouns, each combined with *a* and *the,* and the session when the model first predicted a determiner–noun combination not in the training set up to that session.

homesigners who receive no linguistic input) and compare the models' outcomes to children's actual trajectories.

Finally, we focused here on the development of the determiner category for historical and practical reasons. English determiners have been the focus of much research attention (2–8), likely due to the simplicity of the determiner–noun construction (which involves closed class words, *a* and *the*) and the fact that its acquisition begins relatively early. However, there is nothing in the behavioral and computational framework used here that is specific to this case. The same approach can be applied to constructions involving nouns, verbs, adjectives, adverbs, etc. Comparing the developmental patterns for these constructions in our computational model to the developmental patterns produced by children can help us determine the onset of productivity for the constructions. Assessing the relevant input provided to the model and child can then help us investigate the role that linguistic input plays in determining when productivity for each construction begins.

In sum, we have married longitudinal behavioral observations and computational modeling to capitalize on the strengths of each. Our behavioral data gave us a rich picture of when children begin to productively combine determiners *a* and *the* with the same noun. Although the children produced combinations that were not attested in the sample we had of their parental linguistic input, it is always possible that they had heard these particular combinations at other times in their lives. This is precisely where our computational modeling comes in—we developed a model that did an excellent job of predicting the onsets and trajectories of determiner–noun combinations in the 64 children in our sample. Because we knew exactly what input the model was trained on, we could, with confidence, know that the model had gone beyond the input given. The parallels found between child and model support the claim that the children too were creatively going beyond their input.

Materials and Methods

Corpus Data. The behavioral data for this study come from the LDP (LDP, see ref. 15) corpus, which contained longitudinal observations of 64 typically developing, monolingual, English-learning children from the Greater Chicagoland Area. Children and their primary caregivers were video-recorded, engaging in spontaneous interactions in their homes for twelve 90-min visits (M = 11.3, SD = 1.8, sessions, range 4 to 12 sessions), beginning when children were 14 mo and ending at 58 mo. The resulting corpus of caregiver-child interactions contains over one million transcribed utterances (n = 646,685 for primary caregivers and n = 368,884 for children), and approximately 1,000 h of videos. Both the primary caregiver's and child's utterances were lemmatized, stripped of extraneous punctuation, and all instances of capitalization were removed. All utterances tagged as reading by a human transcriber were excluded for both child and parent. We identified syntactic categories using the part-of-speech taggers provided in the Python spaCy library (26). Additionally, a constituency parse tree was generated for each utterance in the corpus using the Berkeley Neural Parser (27, 28).

Computational Model. A key property of our model is that abstract knowledge of language emerges from domain-general principles of connection-based error-driven learning, based on feedback that is naturally available to children (that is, the actual next word in the data). The focus has been on architectures that can represent utterances as incoming sequences, such as the classic Recurrent Neural Network proposed by Elman (16) and its more recent variants, Long Short-term Memory (29), Gated Recurrent Units (30), and Transformers (14). The latter comes with a technical innovation called self-attention, where the representation of an utterance is built by computing the relationship between each of its words. This mechanism allows the model to exploit distributional information from the input words and dramatically improves performance on a number of tasks in Natural Language Processing. The model known as BERT (31) and its variants, such as RoBERTa (32), have become a focus for computational psycholinguistic research and have been used to simulate many aspects of human language processing, from reading times to brain activities (see ref. 33 for a survey). Although the existing Transformer-based models successfully replicate many empirical patterns observed in humans, they often need much more training data than are available to children. However, a much smaller variation of BERT called BabyBERTa (34) was recently proposed and trained on five million words of data directed at children between the ages of 1 to 6 y. Huebner and colleagues (34) performed postanalysis on the hidden representations of this model and showed that it acquires grammatical knowledge comparable to RoBERTa when pretrained on 160 GB of text.

In our study, we used a variation of BERT, described by Alhama et al. (13). In that study, we ran extensive analyses to validate the model against behavioral data from two different corpora [the Manchester corpus (4), and the LDP corpus used here]. We checked the accuracy of the model in predicting the same determiner that the child produced in the masked slot (*a/an* or *the*, as opposed to predicting any other word), at the end of training (*SI Appendix*, Fig. S2). The model predicts a determiner in 83.43% of the presented utterances, and it predicts the exact determiner in 65.48% utterances.

Experimental Setup. We train our model on child-directed utterances from the corpus under study using the Masked Language Model objective. Since the child-directed data for each individual child are not enough to train the model from scratch, we accumulate utterances of all the parents. We train the model with the input available up to each depicted age.

To test determiner predictions in the model (Studies 1 and 3), we first extract all the determiner usages in the utterances produced by each individual child. Following prior studies (3, 18), we mask the definite and indefinite determiners in DETERMINER + NOUN constructions that follow the pattern DETERMINER + NOUN(SINGULAR) <X> or DETERMINER + <X> + NOUN(SINGULAR), where <X> is any category except NOUN(singular). As an example, for the child's utterance *Here's the pink ear,* we would present the model with *Here's [MASK] pink ear*. We then feed the utterances to the model so that it predicts the most likely filler for the masked slot. For each masked slot, we record the prediction to which the model assigns the highest probability.

To assess the entropy of the model for utterances in which children produced determiners vs. utterances in which children failed to produce a determiner when it was needed, we included all the test utterances in Study 1 and added utterances in which determiners were omitted. Omitted determiners in utterances were identified programmatically using the Berkeley neural parser (28). The final set contained utterances without a determiner but with a noun phrase whose terminal node was a singular common noun (i.e., not a proper noun, nor a pronoun) and did not contain a quantifier ("some," "any"), numeral ("one"), or interrogative ("which"), which could take the place of "a," or "the." The grammaticality of this set of utterances was manually checked by a native speaker of English who had access to the surrounding conversational context. The model was given this set of utterances, where children should have produced a determiner but did not (determiner-omitted), along with utterances from Study 1, where children produced a needed determiner (determiner-produced). The model was asked to fill in a word in the slot where there was a determiner produced or where a determiner should have been produced. The model predicts a probability distribution for the lexical items that could fill in the slot of the missing word. We computed entropy over the predicted probability distribution for both types of utterances (determiner-omitted and determiner-produced).

Data, Materials, and Software Availability. Some study data are available. We will share all code associated with this study, along with in-depth descriptive summaries of the behavioral data underlying the models; however, we do not yet have consent to share the raw child language data previously collected and reported in (15).

ACKNOWLEDGMENTS. This work was supported by NICHD P01 HD 40605 and IES R305A190467 to S.G.-M.

Author affiliations: [a]Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam 1098 XG, The Netherlands; [b]Department of Psychology, New School for Social Research, New York, NY 10003; [c]Department of Psychology, University of Chicago, Chicago, IL 60637; [d]Allen Institute for AI, Seattle, WA 98103; [e]Department of Cognitive Science and AI, Tilburg University, Tilburg 5000 LE, The Netherlands; and [f]Department of Comparative Human Development and Committee on Education, University of Chicago, Chicago, IL 60637

1. L. R. Gleitman, E. Wanner, "Language acquisition: The state of the art" in *Language acquisition: The state of the art*, E. Wanner, L. R. Gleitman, Eds. (Cambridge University Press, New York, 1982).
2. J. M. Pine, E. V. Lieven, Slot and frame patterns and the development of the determiner category. *Appl. Psycholinguist.* **18**, 123–138 (1997).
3. J. M. Pine, D. Freudenthal, G. Krajewski, F. Gobet, Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition* **127**, 345–360 (2013), 10.1016/j.cognition.2013.02.006.
4. A. L. Theakston, E. Lieven, J. M. Pine, C. Rowland, The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *J. Child Lang.* **28**, 127–152 (2001).
5. V. Valian, Syntactic categories in the speech of young children. *Dev. Psychol.* **22**, 562 (1986).
6. V. Valian, "Determiners: An empirical argument for innateness" in *Language down the garden path: The cognitive and biological basis for linguistic structure*, M. Sanz, I. Laka, M. Tanenhaus, Eds. (Oxford University Press, New York, 2013), chap. 14.
7. V. Valian, S. Solt, J. Stewart, Abstract categories or limited-scope formulae: The case of children's determiners. *J. Child Lang.* **36**, 743–778 (2009), 10.1017/S0305000908009082.
8. C. Yang, Ontogeny and phylogeny of language. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6324–6327 (2013).
9. S. C. Meylan, M. C. Frank, B. C. Roy, R. Levy, The emergence of an abstract grammatical category in children's early speech. *Psychol. Sci.* **28**, 181–192 (2017).
10. E. A. Cartmill, D. Hunsicker, S. Goldin-Meadow, Pointing and naming are not redundant: Children use gesture to modify nouns before they modify nouns in speech. *Dev. Psychol.* **50**, 1660 (2014).
11. L. Phillips, N. Hodas, "Assessing the linguistic productivity of unsupervised deep neural networks" in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (London, United Kingdom, 2017), pp. 937–942.
12. B. MacWhinney, *The CHILDES project: Tools for analyzing talk* (Erlbaum Associates, Hillsdale, NJ, 2000).
13. R. G. Alhama *et al.,* "Linguistic productivity: The case of determiners in English" in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, Bali, Indonesia, 2023), pp. 330-343.
14. A. Vaswani *et al.*, Attention is all you need. *Adv. Neural Inf. Process. Syst.* **3**, 6000–6010 (2017).
15. S. Goldin-Meadow *et al.*, New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *Am. Psychol.* **69**, 588 (2014).
16. J. L. Elman, Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
17. J. L. Elman, Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* **7**, 195–225 (1991).
18. C. Yang, "A statistical test for grammar," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (Portland, Oregon, 2011), pp. 30–38.
19. R. Shi, A. Melancon, Syntactic categorization in French-learning infants. *Infancy* **1**, 1–17 (2010), 10.1111/j.1532-7078.2009.00022.x.
20. B. Höhle *et al.*, Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy* **5**, 341–353 (2004).
21. S. Goldin-Meadow, *The resilience of language: What gesture creation* in *deaf children can tell us about how all children learn language* (Psychology Press, New York, 2003).
22. S. Goldin-Meadow, Discovering the biases children bring to language learning. *Perspect. Child Dev.* **14**, 195–201 (2020), 10.1111/cdep.12379.
23. S. Goldin-Meadow, C. Yang, Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neurosci. Biobehav. Rev.* **81**, 150–157 (2017), 10.1016/j.neubiorev.2016.12.016.
24. D. Hunsicker, S. Goldin-Meadow, Hierarchical structure in a self-created communication system: Building nominal constituents in homesign. *Language* **88**, 732–763 (2012).
25. M. Flaherty, D. Hunsicker, S. Goldin-Meadow, Structural biases that children bring to language-learning: A cross-cultural look at gestural input to homesign. *Cognition* **21**, 104608 (2021), 10.1016/j.cognition.2021.104608.
26. M. Honnibal, I. Montani, spaCy2: Natural language understanding with bloom embeddings, convolutional neural networks, and incremental parsing. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 8893–8902 (2017).
27. N. Kitaev, S. Cao, D. Klein, "Multilingual constituency parsing with self-attention and pre-training" in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, Florence, Italy, 2019), pp. 3499–3505.
28. N. Kitaev, D. Klein, "Constituency parsing with a self-attentive encoder" in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Melbourne, Australia, 2018). pp. 2676–2686.
29. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
30. K. Cho, B. V. Merriënboer, D. Bahdanau, Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches" in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 103–111.
31. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, MN, 2019), vol. 1, pp. 4171–4186.
32. Y. Liu *et al.*, RoBERTa: A Robustly optimized BERT pretraining approach. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15846–15851 (2019).
33. A. Karamolegkou, M. Abdou, A. Søgaard, "Mapping brains with language models: A survey" in *Findings of the Association for Computational Linguistics: ACL 2023* (Association for Computational Linguistics, Toronto, Canada, 2023), pp. 9748–9762.
34. P. A. Huebner, E. A. Sulem, C. Fisher, D. Roth, "BabyBERTa: Learning more grammar with small-scale child-directed language" in *Proceedings of the 25th Conference on Computational Natural Language Learning Online* (Association for Computational Linguistics, 2021), pp. 624–646.