Linguistic Productivity: the Case of Determiners in English

Raquel G. Alhama[*]
Tilburg University,[†]
The Netherlands
rgalhama@uvt.nl

Ruthe Foushee[*]
Department of Psychology,
University of Chicago,
USA
foushee@uchicago.edu

Dan Byrne
Department of Psychology,
University of Chicago,
USA
djbyrne@uchicago.edu

Allyson Ettinger
Departments of Linguistics
and Computer Science,
University of Chicago,
USA
aettinger@uchicago.edu

Susan Goldin-Meadow
Departments of Psychology and
Comparative Human Development,
Committee on Education,
University of Chicago,
USA
sgm@uchicago.edu

Afra Alishahi
Tilburg University,
The Netherlands
a.alishahi@uvt.nl

Abstract

Having heard "a pimwit", English-speakers assume that "the pimwit" is also possible. This type of productivity is attributed to syntactic categories such as NOUN and DETERMINER, but the key question is *how* do humans become endowed with these categories in the first place. We propose a novel approach that combines corpus analysis with computational modeling to analyze the productivity of DETERMINER+NOUN constructions in child-produced utterances. Our experiments on two corpora of child-adult interactions using two different methods of quantifying linguistic productivity show that children do not display productivity at early stages. Using a model trained on child-directed utterances, we simulate children's developmental trajectory with great precision, suggesting that the emergence of productivity in human language can be explained without the need to postulate *a priori* access to syntactic categories.

1 Introduction

Having heard "a pimwit," English-speakers know immediately that "the pimwit" is possible, even if they have not heard the phrase before. Researchers from diverse theoretical perspectives agree that this type of productivity can be explained with syntactic categories (in this case, DETERMINER and NOUN), but the key question is *how* do humans become endowed with these categories in the first place. The acquisition of the English definite and indefinite determiners (*the*, *a*) has been frequently used as a case study to tackle the question of whether syntactic categories are present at birth, or rather are learnt from exposure to linguistic data.

Linguistic productivity provides evidence for syntactic categories. If English-learning children have a DETERMINER category, they should be able to produce the same noun with different determiners (i.e., a child would produce both *a dog* and *the dog*) (Pine et al., 2013; Yang, 2011). If they do not have a DETERMINER category, a child who hears *a dog* may not immediately understand that *the dog* is also possible. On this reasoning, when a child begins to use the same noun with both *a* and *the*, it can be seen as evidence that the child possesses a productive determiner category.

Much effort is put into formalizing a quantitative measure of productivity that can be applied to spontaneous child productions and give a reliable estimate of *when* young language learners start to use syntactic categories productively in their speech. Pine and Lieven (1997) propose such a measure, *overlap score*, which estimates determiner productivity based on the proportion of nouns in child-produced language that are used with both determiners. They estimate determiner productivity for children in the Manchester corpus (**?**) and argue that children are not fully productive at early stages of learning. Yang (2011) questions the validity of the original *overlap score* and proposes a revised version which takes the Zipfian distribution of nouns and determiners into account. Their analysis of the new score on Manchester corpus suggests that determiner productivity in children is not quantitatively different from that of adults. However, Pine et al. (2013) argue that Zipfian distributions

[*] Equal contribution.

[†] The author is currently affiliated with the Institute for Logic, Language and Computation, University of Amsterdam.

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 (Volume 1: Long Papers), pages 330–343
November 1–4, 2023. ©2023 Association for Computational Linguistics

are not a good approximation of noun frequency in children's productions. Additionally, Meylan et al. (2017) simulate determiner productivity in a Bayesian model and show that the previous analyses use input data that is too small to yield reliable conclusions. In addition to the size of the input corpus, Meylan et al. (2017) highlight another important limitation of using the *overlap score*. This measure is quite sensitive to the amount of child-produced data, and therefore cannot be reliably used to estimate productivity at the very early stages of learning (when children do not produce many DETERMINER+NOUN combinations). Their simulation results suggest that the youngest age groups in previous behavioural studies might not be young enough for their data to show productivity. This paper makes the following contributions. We propose a novel approach that combines corpus analysis with computational modeling. We use a data-driven computational model with no built-in categories to simulate the process of language learning from child-directed utterances. To address the limitations related to the input corpus, in addition to reproducing behavioral patterns of previous studies on the Manchester corpus, we also use the Language Development Project (LDP) corpus (Goldin-Meadow et al., 2014), which contains data from many more children and records interactions from a much younger age. To address the concerns regarding the productivity measure, we estimate DETERMINER+NOUN productivity in the utterances produced by humans and by the model not only using the *overlap score* but also the *onset measure*, a simple and data-efficient estimate of productivity (Cartmill et al., 2014). Altogether, this approach allows us to examine the developmental trajectory of the DETERMINER+NOUN construction in children, and the extent to which it can be explained by a model that has no prior access to syntactic categories. We show that the behaviour of our computational model closely mimics the patterns observed in children over time, revealing the onset of determiner productivity in both model and child. These results strongly suggest that children's linguistic productivity can be achieved based on learning from statistics of the child-directed data.

2 Related Work

Many computational models have framed the general problem of inducing abstract categories from unannotated text as clustering words into lexical categories based on the distributional properties of their context (e.g., Redington et al., 1998; Clark, 2000; Mintz, 2003; Parisien et al., 2008; Chrupała, 2011), showing the possibility of learning categories that resemble parts of speech from raw text. Alishahi and Chrupała (2012) and Abend et al. (2017) model concurrent acquisition of word meanings and syntactic categories and focus on the impact of integrating knowledge of syntax (and particularly the syntactic category of a word) into the word learning process. However, none of these models focus on the nature and developmental trajectory of the induced categories, nor do they compare their linguistic productivity to humans. An exception is Parisien et al. (2008) who present an incremental Bayesian model for learning syntactic categories from linguistic context, and test it on child-directed data from the Manchester corpus. Their analysis of the emerging categories shows that the categories follow the same trend as children's categories in that nouns are learned before verbs, followed by adjectives (Kemp et al., 2005). However they do not analyze the emergence of each category and its use in child-produced speech.

With the increased dominance of deep neural models of language, much effort is put into analyzing the learned representations in inner layers of these models and to search for encoding syntactic information (see Belinkov and Glass, 2019, for an overview). Various analyses have shown that deep language models encode information about syntactic categories and syntactic dependencies in their learned representations without explicit training (e.g., Adi et al., 2017; Hewitt and Manning, 2019; Chrupała and Alishahi, 2019; Tayyar Madabushi et al., 2022). However, the focus of this body of work is mainly on large language models that are trained on massive datasets (with some exceptions; see Grimm et al., 2015), and comparison with human language learning is not common. Pannitto and Herbelot (2020) and Huebner et al. (2021) are two exceptions, where they each train a neural network from scratch on child-directed utterances and compare the output of the model to human productions. Although these studies are not specifically focused on the emergence of syntactic categories, we use the latter as an inspiration for our own modeling experiments.

Computational studies of the acquisition of the determiner category are scarce. One such study is the above-mentioned work by Meylan et al. (2017), which used a hierarchical Bayesian model for simulating syntactic productivity in the case of DETERMINER+NOUN constructions, with parameters to represent the role of experience and an *a priori* tendency to generalize. Their experiments suggest that previous corpus studies use input corpora that are too small to yield statistically reliable conclusions. They used a large dataset of child-adult interactions for a single child and found determiner-noun productivity, but only if the parameters of the model are set in a specific way to encourage generalization (as opposed to memorization). However, the Meylan et al.'s model works best when applied to a large sample of data from each child, and is not applicable in early stages of learning when child-produced language is scarce.

One other study that uses the overlap score to measure determiner productivity in a deep neural network model is Phillips and Hodas (2017). This model uses an autoencoder architecture whose objective is to reconstruct (or repeat) an input utterance, and train it on child-directed utterances from a collection of corpora from CHILDES (MacWhinney, 1995). The model learns a compact, latent representation for every incoming utterance, which is then used to regenerate the same utterance. They measure the estimated and empirical overlap scores in adult utterances from the training corpora and in the utterances generated by the autoencoder model, and show that if the model's parameters are set to allow for more generalizability, its estimated overlap scores are closer to those of humans. This study does not compare the behaviour of the trained model to the behavior of language learning children, and therefore says little about the learning trajectory of the determiner class.

3 Method

In the following sections, we present a series of interleaved computational experiments and behavioral analyses to examine linguistic productivity.

3.1 Data

The Manchester corpus This corpus records a study of 12 monolingual English-speaking children from middle-class households in Manchester, UK, from ages 20 to 36 months. Mothers and children were audio-recorded playing freely in their homes two times every three weeks for a year, for a maximum of 34 sessions per child. At the beginning of the study, the children ranged in age from 1;8.22 to 2;0.25 with mean length of utterances (MLUs) ranging between 1.06 to 2.27 in morphemes (Theakston et al., 2001).

The LDP corpus The Language Development Project corpus (LDP) followed 64 typically developing, monolingual, English-speaking children from the Greater Chicagoland Area. Children and their primary caregivers were video-recorded engaging in spontaneous interactions in their homes for twelve 90-minute visits (M=11.3, SD = 1.8, sessions, range 4–12 sessions), beginning from when the children were 14 months to 58 months. The resulting corpus of caregiver-child interactions contains over 1 million transcribed utterances ($n = 646,685$ for primary caregivers and $n = 368,884$ for children), and approximately 1,000 hours of videos (Goldin-Meadow et al., 2014).

Preprocessing Both the primary caregivers' and children's utterances were lemmatized, stripped of extraneous punctuation, and all instances of capitalization were removed. All utterances tagged as reading aloud were excluded. We identified syntactic categories using the part-of-speech taggers provided in the spaCy library (Honnibal and Montani, 2017). Our preprocessing pipeline is shared online[1].

3.2 Productivity Measures

We consider two methods of assessing grammatical productivity in our behavioral as well as model-generated data.

Overlap score Following Pine et al. (2013), we compute the overlap score as the number of noun types that the child pairs with both *a* and *the*, divided by the total number of noun types used with determiners. We refer to this metric as 'observed overlap', to distinguish it from the 'expected' overlap variant proposed in Yang (2011).[2] Since this metric is very sensitive to sample size (Meylan et al., 2017), we only report the score for sample sizes that consist of at least 50 productions.[3]

Onset measure We assume that a child has productive use of the category under study when the child uses all of its forms (in our case, *a* and *the*)

[1] https://osf.io/s2jnm/?view_only=ca2d57aee759426ba1c531a64bc982f0

[2] We ran all of our analyses also with the 'expected' overlap metric and found a high correlation between both metrics.

[3] This number was chosen as a threshold after personal communication with Charles Yang.

with the same noun, and does so with at least two different nouns within the same session (i.e., *a car, the car, a bottle, the bottle*). This criterion is inspired by Cartmill et al. (2014); the only difference is that we do not require that the child continue using the category in the upcoming sessions.

4 Computational Model

Our goal is to examine the trajectory of determiner productivity in a data-driven computational model trained on child-directed data, and see to what extent it resembles the same trajectory in children. Although we do not seek to simulate the exact mechanism that children use for learning and processing language, it is important to choose a modeling framework and architecture that satisfies a number of criteria. First, the model must not rely on any data or supervision signal that is not available in children. Second, the ultimate task on which the model is trained must be similar to the daily experience of children using language. Third, the model must not rely on any explicit, latent representation of abstract syntactic categories in advance. A model that meets these requirements would allow us to study this phenomenon at Marr's computational level (Marr, 2010).

Many computational models that simulate syntactic category acquisition from large corpora falsify the first criterion, for example by relying on corrective feedback on part of speech labels assigned to words, or on which words must be clustered together (see Section 2 for an overview). Similarly, a number of existing models that are used for investigating the current debate on determiner productivity do not adhere to the second criterion; for example Phillips and Hodas (2017) use an autoencoder architecture trained on a repetition task, where a model learns to regenerate input utterances, which is an unrealistic task from a language learner's point of view. Criterion three is the core of our study: we want to see how far we can go in reproducing productivity patterns in children without assuming pre-existing abstract categories.

A natural choice of model for our enterprise is Transformers (Vaswani et al., 2017). First of all, training Transformers with the parental input and the Masked Language Modelling (MLM) objective is adequate in terms of the information available in the supervision signal, which consists uniquely of information available to children (words in caregiver's utterances). Second, we can use the MLM objective to test the model by masking the determiners in children's productions. In this way, we can simulate children's productivity, thereby reproducing a more naturalistic scenario than previous approaches. Finally, this model does not rely on any pre-existing syntactic knowledge, which meets our third objective.

The existing Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become a focus for computational psycholinguistic research and used to simulate many aspects of human language processing, from reading times to brain activities (see Frank et al., 2019, for an overview), but they often need much more training data than is available to human language learners. However, a much smaller variation of BERT called BabyBERTa (Huebner et al., 2021) was recently proposed and trained on 5 million tokens of data directed at children between ages of one to six. They performed post-analysis on the hidden representations of this model and showed that it acquires grammatical knowledge comparable to RoBERTa pre-trained on 160GB of text. We follow this approach and use a much smaller variation of BERT in our study which we train from scratch on child-directed data.

We instantiate a small Transformer-based model using the HuggingFace library (Wolf et al., 2020) and reduce the number of attention heads to 2 and the number of hidden layers to 1. We tokenize the input with WordPieceTokenizer (Wu et al., 2016), with a maximum vocabulary size of 30000 and a minimum frequency of 2. We use the default training configuration provided in HuggingFace's Trainer (Adam), with the exception of the batch size, which we increase from 8 to 64. Remaining hyperparameters are set to their default values in HuggingFace (see exact version in our shared code[4]). Crucially, we do not apply any pre-training on this model; therefore, our initial model does not have any built-in linguistic knowledge.

4.1 Experimental Setup

We train our model on child-directed utterances from the corpus under study, using the MLM objective. Since the child-directed data for each individual child is not enough to train the model from scratch, we accumulate utterances of all the parents. In experiments showing a developmental trajectory over time, we train the model with the input avail-

[4] https://github.com/rgalhama/defdets_aacl

able up to each depicted age.

To test the determiner productivity of the model, we first extract all the determiner usages in the utterances produced by each individual child. Following prior studies (Pine et al., 2013; Yang, 2011), we mask the definite and indefinite determiners in DETERMINER+NOUN constructions; in particular "DETERMINER NOUN-SINGULAR <X>" and "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.

4.2 Overall performance

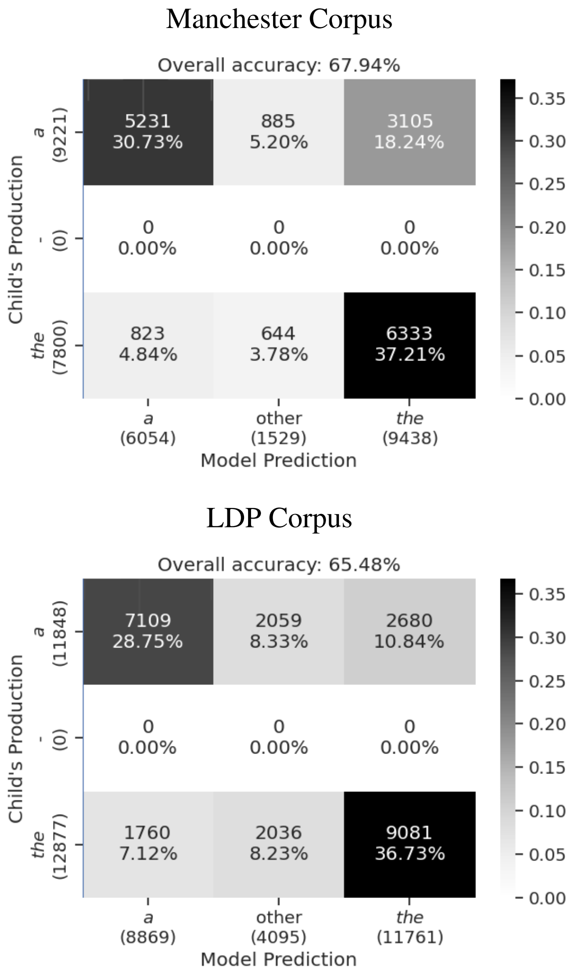


Figure 1: Confidence Matrix of masked determiner predictions for the model trained on all sessions in Manchester corpus (top) and in LDP corpus (bottom).

To make sure that the model is properly trained, we checked the accuracy of the model in predicting a determiner in the masked slot.We compute the accuracy of the predictions for the sentences produced by children in a certain session (age) with the model trained with input data up to that session. Figure 1 reports the accuracy of predicting the masked determiner in children's productions by the models trained on data from all sessions. In the case of the Manchester corpus, the model is 67.94% accurate in predicting the exact determiner, while it predicts any of the two determiners with a probability of 91%. For the LDP corpus, the model can predict the exact determiner with similar performance (65.48%) . The accuracy for predicting either of these two determiners is 83%.

Both models are more likely to predict the definite determiner (*the*) over the indefinite (*a*). This is more noticeable in the case of the Manchester corpus, in which 55.45% of the predictions are for the definite determiner (47.13% for the LDP data). This likely reflects the fact that *the* is more frequent in parental speech. For instance, in the case of parental data in LPD, 61% of the constructions that we study (which combine definite and indefinite determinants with singular nouns, as explained in section 4.1) use *the*, versus 39% using *a* (but note that *the* can also appear in the input combined with plural nouns). Despite this bias in the data, the accuracy of the model in predicting the exact determiner is above chance for both models. (The average probability mass of the decoded determiners can be found in Appendix C.)

Child	**Model**
You're the monster	You're a monster
I see a moon	I see the moon
Up a stair	Up the stair
Is this a bug?	Is this the bug?
Draw a train track	Draw the train track
Tickle the back	Tickle a back
No, I want the penguin story	No, I want a penguin story
He take a bathroom	He take the bathroom

Table 1: Some examples where the model produces a different determiner from the child.

Note that the overall performance of the model is underestimated as some of the model's predicted determiners that mismatch the ones used by the child (and therefore considered as errors) might be plausible choices in the given context. Table 1 shows a few examples of such errors, some of which sound more plausible than the child's choice (e.g. *You're a monster*).

5 Our Study of Linguistic Productivity

To investigate the development of linguistic productivity in children and to simulate the process using our computational model, we conduct the following three experiments in this study.

Experiment 1. We establish the reliability of our model by reproducing the behavioral patterns observed in Pine et al. (2013) and Yang (2011): we train our model on child-directed utterances from the Manchester corpus, and use the trained model to predict the masked determiners in child-produced sentences in the same corpus. Following those studies, we estimate and compare overlap scores in the utterances produced by adults, children and our model.

Experiment 2. We train the model on child-directed utterances from the LDP corpus, which contains data from more children and records interactions from a younger age. We then compare the estimated overlap scores in the utterances produced by adults, children and our model.

Experiment 3. We use the same model trained on child-directed utterances from the LDP corpus, but this time we compare the patterns of production of DETERMINER+NOUN combinations by adults, children and by our model using the onset measure.

Through these experiments, we show that our model closely mimics the behaviour of children when using determiners, but the use of LDP corpus and the choice of the onset measure presents a clearer picture of emergence of determiners as a syntactic class.

5.1 Experiment 1: Overlap Score on Manchester Corpus

To reproduce previous studies, we first train our model on child-directed utterances from all parents in the Manchester corpus. After the model is fully trained, for each of the 12 children in this corpus we take all the sentences that contain a DETERMINER+NOUN usage, mask the determiner and feed the masked input to the model, then replace the masked word with the top determiner predicted by the model (see Section 4 for details).

Figure 2 shows the observed overlap scores as measured for the sentences produced by the Manchester parents, children, and our model. We see that there is no noticeable difference between overlap scores of children and adults, which is in line with what Yang (2011, 2013) reports. However, the overlap scores estimated for utterances produced by our model are closer to those of children than to adults.

Due to the small size of the Manchester corpus, we could not train the model on this data incrementally and trace the trajectory of overlap scores as a function of the age of children in this corpus. We address this issue in the next experiment.

5.2 Experiment 2: Overlap Score on LDP Corpus

Although Yang (2011) interprets similarity between overlap scores for children and adults as evidence for the availability of *a priori* syntactic categories, children participating in the earliest sessions of the Manchester corpus might already be old enough to have acquired the abstract determiner class. In addition, the sample of children (n=12) is quite small (Meylan et al., 2017). Therefore we perform the same experiment on the LDP corpus.

We train the model on child-directed utterances from the LDP corpus and use it to predict masked determiners in child-produced utterances. Figure 3 shows overlap scores for utterances produced by LDP parents, children and our model, where children (on the x axis) are ranked by their estimated overlap scores. As before, the overlap scores for all three groups are close to each other, with those for the model closer to children than adults. There is a stronger fit to the predictions of our model and the LDP children compared to those of the model and adults. To see whether the distance between the overlap score for children and adults changes as children grow older, we also estimated overlap scores averaged over all children within the same session; that is, we accumulated all the utterances that parents and children were producing at a given point in time. This time the pattern is different: as can be seen in Figure 4, there is a gap between overlap scores for parents and children in the earlier sessions, which disappears in the later sessions. However, due to the small number of utterances produced by children at earlier sessions, the estimation of overlap score is not reliable (in fact we could not apply the measurement to earlier ages because the sample size was smaller than 50; see Section 3.2). As before, overlap scores estimated for the model are closer to children than adults.

5.3 Experiment 3: Onset on LDP Corpus

To address the limitations of the overlap score and get a better picture of determiner productivity in

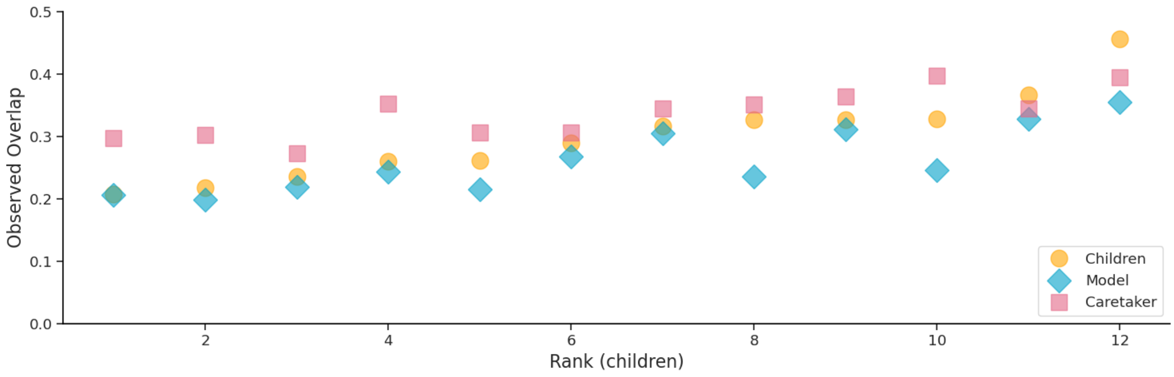


Figure 2: Overlap over Manchester corpus. The x-axis is sorted by observed overlap in child-produced language.

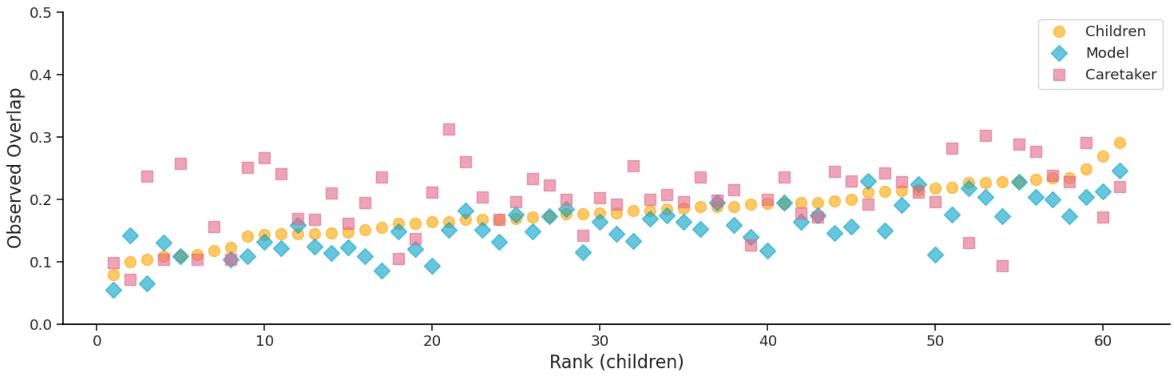


Figure 3: Observed overlap over LDP corpus.

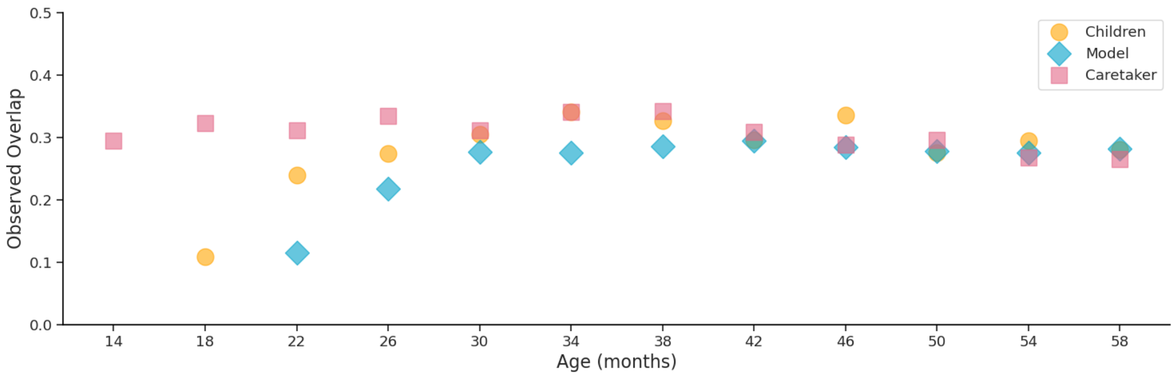


Figure 4: Observed overlap over LDP corpus, averaged by age group.

early stages of development, we repeat the previous experiment on LDP corpus but this time we use the onset measure to estimate linguistic productivity.

Figure 5 shows the median number of noun types that appear with both determiners in each session; the dashed black line corresponds to the onset criterion (i.e., when a minimum of two noun types has been used with both determiners in the same session). The points above this line correspond to learners who have achieved determiner productivity. We see that while productivity is relatively constant in parental speech, children use determiners for several months without displaying productivity. This onset pattern is replicated in the model. As in the previous experiments, we can see that the model shows an excellent fit to the trajectory of children. The small decrease in productivity at the older ages is due to the fact that both children and parents produced fewer utterances overall at these points. In Appendix D we have included a

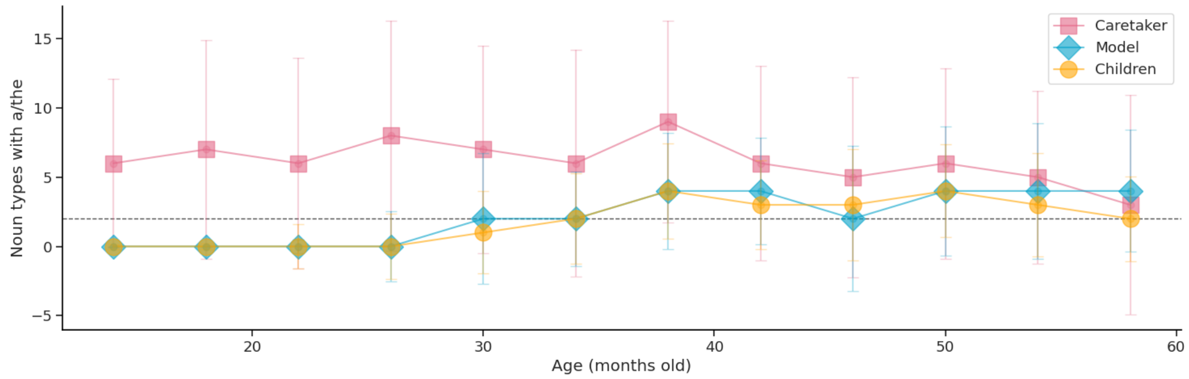


Figure 5: Median number of noun types that are produced with both determiners. The dashed horizontal line represents the lower boundary for our onset criterion (i.e., when two different nouns are each produced with both determiners). Error bars correspond to standard deviation.

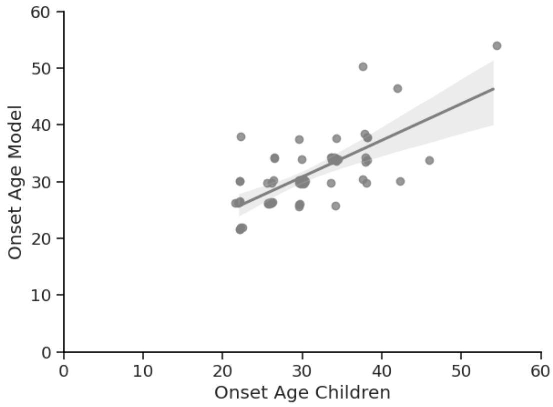


Figure 6: Age of productivity according to the onset measure for LDP children and the model ($r = 0.71$). A random jitter of 0.5 has been added to overlapping points in the graph.

graphical demonstration of this pattern for each individual child in LDP to show that this fit is not a result of averaging over all children.

In addition to producing more noun types with both determiners over time, the moment of onset of determiner productivity in the model also shows a strong correlation with the moment of onset in children. Figure 6 shows this correlation for all the children in LDP and their corresponding model simulation.

It is worth mentioning that the onset measure also depends on the amount of input produced by children, as there is a correlation between the median number of words produced by children and the number of productive noun types in each session (Kendall's rank correlation τ = -0.25, p=.007); however, we found that 84% of the children had already produced two exemplars of two different nouns in at least one session prior to the hypothesized onset, but without passing the onset criterion. In other words, the children could have displayed productivity, but did not.

Impact of linguistic context. A possible concern could be that the similarity between the predictions of the model and children might be due to properties of child-produced utterances, rather than the internal representations of syntactic categories. To investigate this potential confound, we ran a control experiment in which we tested the model also on utterances that the parents produced (see Appendix E for details). We found no significant difference between model predictions on parent-produced versus child-produced utterances, suggesting that our findings are not driven by the context used for the test.

Productivity vs. memorizing. Given the goodness of fit of the model, there is the possibility that the model simply memorizes DETERMINER+NOUN usages it has seen in the training data instead of productively using abstract categories. Since (unlike children) we have full access to the input of the model, we can search for the novel combinations in its output. Table 2 shows some examples where the model produces a determiner it has never seen before in combination with the target noun, but is more appropriate for the context. This is an important observation which hints at the abstract nature of the emerging syntactic categories.

6 Conclusion

In this paper, we investigate *how* and *when* English-speaking children productively use DETER-

Age (m.o.)	Model Prediction	Noun Frequency
22	Take *a milkie*	9
26	Sylva has *a flu*	5
30	That's *a daffodil*	3
42	You just call him *an idiot*	0
46	Do you want to have *a boyfriend?*	5

Table 2: Examples of model-produced DETERMINER+NOUN combinations that were not in the training data. Age is in months old.

MINER+NOUN constructions. Our experiments on two corpora of child-adult interactions suggest that children do not use determiners productively when they first produce them. It isn't until 30 months that children display productivity in their use of determiners. Thanks to the use of the data-efficient onset measure, this outcome is less dependent on sampling effects and therefore more reliable.

We furthermore train a Transformer-based model from scratch on child-directed utterances, and simulate determiner production in children. Our model mimics children's developmental trajectory with great precision, suggesting that the emergence of productivity in human language can be explained without the need to postulate access to pre-existing syntactic categories.

Our hybrid approach is independent of the target syntactic construction. In the future, we plan to apply this methodology to a range of syntactic phenomena, and investigate the trajectory as well as the order of development of different syntactic categories in children exposed to different languages.

Limitations

In this study we combined child-directed data from all caregivers in our corpus to train the computational model, and only used the individual child-produced utterances as test cases for each version of the model. This was due to lack of enough training data, but in the future we must think about creative ways of simulating individual differences (for example by pre-training the model on a generic child-directed corpus, and fine-tune each instance of the model on child-directed data from individual children in our corpus).

In addition, we focused on the development of only one syntactic category, and only in English. This is due to historical and practical reasons: English determiners have been used as a case study in this domain, likely due to the simplicity of the DETERMINER+NOUN construction (which involves closed vocabulary items rather than an open lexical class) and the fact that its acquisition seems to start relatively early. However, there is nothing in the empirical and computational framework we used here that is specific to this particular case, and the same approach can be applied to any other languages and prominent categories such as nouns, verbs, adjectives and adverbs. Comparing their developmental pattern in the computational model can allow us to investigate the order of development often hypothesized in children (e.g. nouns are learned before verbs, adjectives become productive much later, etc.).

Ethics Statement

We used the LDP dataset Goldin-Meadow et al., 2014, a corpus that was collected from spontaneous parent-child interactions recorded in a home-setting. This rather personal, intimate setting lends itself to the frequent use of Personally Identifiable Information (PII), either directly in utterances or indirectly through personal artifacts.

The data collection protocol (approved by University of Chicago, IRB protocols 02-942 and H12078) determined the following privacy considerations:

1. Subjects were issued subject numbers.
2. Paper files were identified by number only and kept in locked file drawers.
3. Electronic files were stored on password-protected computers that were accessible only research team members. Other researchers only had access to coded data.
4. Videos were stored on a securely managed, password-protected server and each researcher was given his/her own secure login and password. Any local copies of videos were deleted.
5. Videos were not linked to any other identifying information except what is contained in the video.
6. During transcription, identifiers were removed.
7. The master list was stored in a password-protected computer, was kept separate from data and was only kept during data collection.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Afra Alishahi and Grzegorz Chrupała. 2012. Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 643–654.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Erica A Cartmill, Dea Hunsicker, and Susan Goldin-Meadow. 2014. Pointing and naming are not redundant: children use gesture to modify nouns before they modify nouns in speech. *Developmental Psychology*, 50(6):1660.

G. Chrupała. 2011. Efficient induction of probabilistic word classes with LDA. In *International Joint Conference on Natural Language Processing*.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

A. Clark. 2000. Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2nd workshop on Learning Language in Logic and the 4th conference on Computational Natural Language Learning*, pages 91–94. Association for Computational Linguistics Morristown, NJ, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan L Frank, Padraic Monaghan, and Chara Tsoukala. 2019. Neural network models of language acquisition and processing. In *Human language: From genes and brain to behavior*, pages 277–293. MIT Press.

Susan Goldin-Meadow, Susan C Levine, Larry V Hedges, Janellen Huttenlocher, Stephen W Raudenbush, and Steven L Small. 2014. New evidence about language and cognitive development based on a longitudinal study: hypotheses for intervention. *American Psychologist*, 69(6):588.

Robert Grimm, Giovanni Cassani, Walter Daelemans, and Steven Gillis. 2015. Towards a model of prediction-based syntactic category acquisition: First steps with word embeddings. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 28–32, Lisbon, Portugal. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Nenagh Kemp, Elena Lieven, and Michael Tomasello. 2005. Young children's knowledge of the" determiner" and" adjective" categories. *Journal of Speech, Language & Hearing Research*, 48(3).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

B. MacWhinney. 1995. *The CHILDES Project: Tools for Analyzing Talk*, second edition. Hillsdale, NJ: Lawrence Erlbaum Associates.

David Marr. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Stephan C Meylan, Michael C Frank, Brandon C Roy, and Roger Levy. 2017. The emergence of an abstract grammatical category in children's early speech. *Psychological science*, 28(2):181–192.

T.H. Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online. Association for Computational Linguistics.

Christopher Parisien, Afsaneh Fazly, and Suzanne Stevenson. 2008. An incremental bayesian model for learning syntactic categories. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 89–96.

Lauren Phillips and Nathan Hodas. 2017. Assessing the linguistic productivity of unsupervised deep neural networks. In *Proceedings of the 39th Annual Meeting of the Cognitive Sciencey Society (CogSci 2017)*, pages 937–942, London, United Kingdom. Austin, Texas:Cognitive Science Society.

Julian M. Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. 2013. Do young children have adult-like syntactic categories? zipf's law and the case of the determiner. *Cognition*, 127(3):345–360.

Julian M Pine and Elena VM Lieven. 1997. Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(2):123–138.

M. Redington, N. Crater, and S. Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science: A Multidisciplinary Journal*, 22(4):425–469.

Harish Tayyar Madabushi, Dagmar Divjak, and Petar Milin. 2022. Abstraction not memory: BERT and the English article system. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 924–931, Seattle, United States. Association for Computational Linguistics.

Anna L. Theakston, Elena Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28:127–152.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Charles Yang. 2011. A statistical test for grammar. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–38, Portland, Oregon, USA. Association for Computational Linguistics.

Charles Yang. 2013. Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327.

A The LDP Corpus: Recruitment and Demographics

The data collection of the LDP corpus is fully described in Goldin-Meadow et al. (2014). To recruit participants, advertisements were posted in the Chicago Parent magazine, a recruitment letter was written and sent out, flyers were posted, and daycare centers were contacted. 50$ were given per visit to parents for an approximately 2-hour-long visit. Parents who responded participated in a screening questionnaire over the phone during which information was gathered on ethnicity, income, education, language(s) spoken in the home, and child gender.

Sixty-four English-speaking families were selected to match as closely as possible the 2000 census data on family income and ethnicity in the greater Chicago area.

B Computational Simulations

In this section we report additional details of the computational simulations. The number of parameters of our model is much smaller than that of the original BERT model (16960546 for the Manchester data, and 17373499 for LDP[5]). We set the hyperparameters of the models (specified in section 4.1) manually, after a few trials. Given the smaller size of models and data, they can even be trained on a laptop with a modest CUDA-compatible GPU (training time is shorter than 8 hours in an MSI Prestige A10SC-006NL with an NVIDIA GeForce GTX 1650).

The training data for the Manchester corpus consists of 351223 child-directed sentences and 1767057 word tokens. In the case of the LDP corpus, the full size of the training data (also consisting of child-directed sentences) is 646040 sentences and 2544468 word tokens. Table 3 reports the training data available to the model at each age, for the version of the model trained incrementally.

To test the model, we use child-produced (rather than child directed) sentences, hence the training and test data do not have any overlap. In the case of the Manchester corpus, the number of sentences used to test the model are enumerated in Table 4, separated by individual child. The equivalent information for the LDP corpus is summarized in Figure 7 (given the greater number of children, this more compact visualization is more convenient than a table). The size of the test data used to analyze the model by age is presented in Table 5.

Age	Sentences	Words (tokens)
14	63649	214612
18	64660	219182
22	60574	215085
26	59538	225807
30	57358	225877
34	53787	217869
38	59619	239189
42	53237	215223
46	47364	196980
50	46950	208677
54	39555	175709
58	41128	190258

Table 3: Training data size for LDP, by children's age.

Child	DETERMINER+NOUN Sentences
Anne	1133
Aran	1731
Becky	1340
Carl	3627
Dominic	389
Gail	848
Joel	1080
John	1755
Liz	1392
Nicole	690
Ruth	792
Warren	2256

Table 4: Test data size of Manchester corpus, by child.

Age	DETERMINER+NOUN Sentences
14	51
18	414
22	850
26	2062
30	3073
34	3546
38	4388
42	4010
46	4576
50	4468
54	3781
58	3973

Table 5: Test data size of LDP corpus, by children's age.

C Average Maximum Probability

Our model provides a probability distribution over the vocabulary items, which is then used to predict the word that should fill in the masked slot. As a decoding strategy, we choose the word with maximum probability (the mode). To make sure that this is not an ill-founded choice (as the probability

[5] Note that parameters vary with vocabulary size.

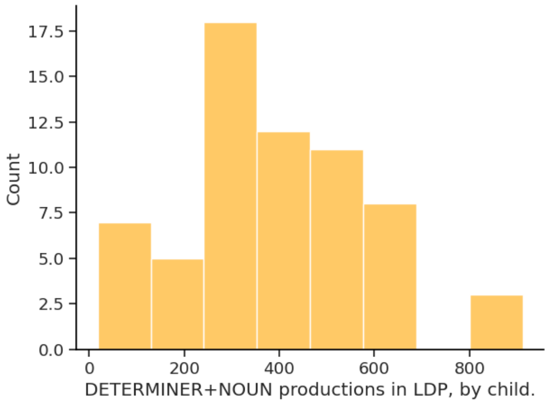


Figure 7: Histogram of the size of the test data by individual child in the LDP corpus, for n=63 children (the data of 64th child was deemed unusable due to not having enough observations).

mass attributed to the mode may still be a small percentage of the overall mass), we report the average probability mass that the model attributed to the predicted words.

For the model trained on the Manchester corpus, the average probability mass for the preferred determiner (for predictions of the exact same determiner) was 0.41±0.24. This is quite substantial, considering that the model needs to distribute this probability between all the units in its vocabulary. If we consider all the sentences for which the model predicted a determiner, then the average probability mass of the top prediction is 0.37±0.23. The average maximum probability across all the tested sentences is 0.24 ±0.19.

A similar pattern is observed for LDP, although in this case the model has higher certainty, likely because of the greater amount of training data. When trained on this dataset, the average probability mass for the preferred determiner (for predictions of the exact same determiner) was 0.52±0.21. When considering all the sentences for which the model predicted a determiner, then the average probability mass of the top prediction is 0.49±0.21. The average maximum probability across all the tested sentences is 0.46 ±0.22.

D Onset measure applied to individual children in LDP

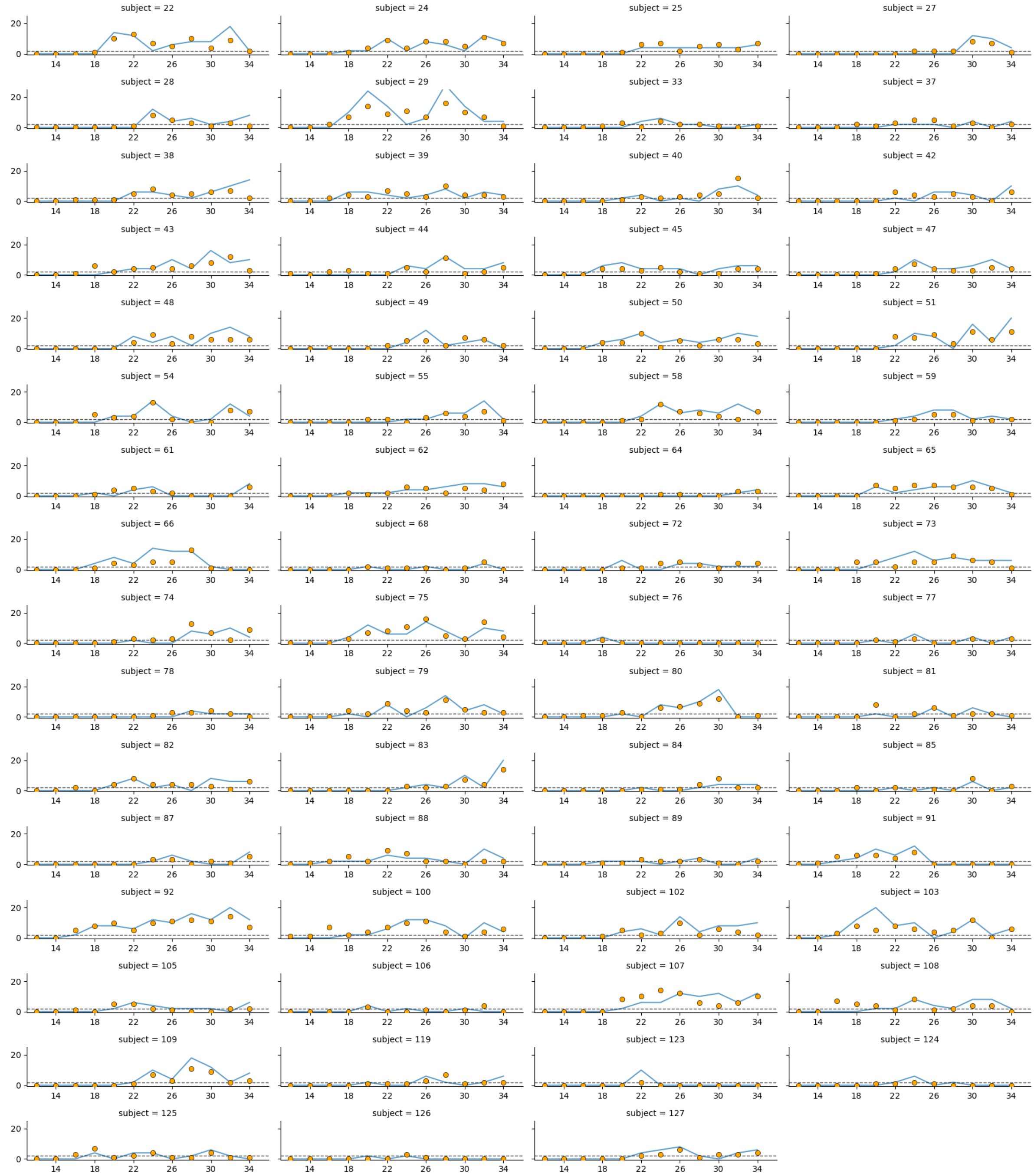
Figure 8 shows the Onset metric applied to LDP data, for children and model predictions. Each subgraph corresponds to an individual child and its corresponding model (i.e. a model tested with the utterances produced by the individual child).

	t-statistic	p-value
1	-1.633	0.178
2	0.552	0.601
3	1.000	0.374
4	0.111	0.919
5	-1.400	0.220
6	-1.581	0.175
7	-1.718	0.161
8	0.542	0.611
9	0.535	0.621
10	-0.159	0.880

Table 6: Results of significance test, applied over model predictions on child-produced and child-directed speech.

E Control Experiment

To control for the possibility that the good fit of our model to the children data was due to testing the model using child-produced speech, we ran the following experiment: for each session, we sampled utterances with determiner constructions (n=48) separately for child utterances and for parent utterances, and used them as test frames for the model. We computed the productivity of our model for each set of sentences and performed a paired statistical test. We repeated this experiment 10 times, and none of these tests yielded a significant difference between predictions for determiner-noun test frames taken from parent speech and predictions for determiner-noun test frames taken from child speech. This result rules out the hypothesis that our findings are due to the linguistic context of child-produced speech. The results can be found in Table 6.



subject = 22
subject = 24
subject = 25
subject = 27
subject = 28
subject = 29
subject = 33
subject = 37
subject = 38
subject = 39
subject = 40
subject = 42
subject = 43
subject = 44
subject = 45
subject = 47
subject = 48
subject = 49
subject = 50
subject = 51
subject = 54
subject = 55
subject = 58
subject = 59
subject = 61
subject = 62
subject = 64
subject = 65
subject = 66
subject = 68
subject = 72
subject = 73
subject = 74
subject = 75
subject = 76
subject = 77
subject = 78
subject = 79
subject = 80
subject = 81
subject = 82
subject = 83
subject = 84
subject = 85
subject = 87
subject = 88
subject = 89
subject = 91
subject = 92
subject = 100
subject = 102
subject = 103
subject = 105
subject = 106
subject = 107
subject = 108
subject = 109
subject = 119
subject = 123
subject = 124
subject = 125
subject = 126
subject = 127

Figure 8: Number of noun types that appear with both determiners, for individual children (orange dots) and corresponding model (blue line). Dashed line corresponds to lower boundary to achieve our criterion for productivity (2 nouns with 2 determiners). Horizontal axis represents age, in months old.