# Generalizing Without Evidence: How Transformer Models Infer Syntactic Rules From Sparse Input

Mark van den Hoorn

Faculteit der Geesteswetenschappen, Universiteit van Amsterdam

139229002Y: Bacheloronderzoek Cognition, Language and Communication

Dr Raquel Garrido Alhama

24 June 2025

**Abstract**

The debate between generativist and constructivist theories of language acquisition centers on whether children possess innate syntactic rules or learn them from experience with linguistic input. While computational models have been useful to simulate the process of language learning, it remains unclear what specific input features are necessary for grammatical generalization. This thesis addresses that gap by focusing on English determiners, asking whether a Transformer-based model (specifically BERT) can generalize determiner use when trained on data where determiner-noun variability is artificially limited. To test this, a set of BERT models were trained from scratch on child-directed speech, with five input conditions ranging from natural to completely restricted, where nouns were paired with only ‘a’ or only ‘the’, but never both. The models' ability to generalize was evaluated on their predictions of masked determiners in child-produced sentences. The results show that all models, including the one trained on the most restricted input, successfully generalized by productively using determiners in combinations they had never seen. While all models learned the abstract 'determiner' category, accuracy for predicting the specific masked determiner decreased as input restrictions increased. These findings suggest that direct exposure to determiner-noun variability is not a necessity for acquiring an abstract grammatical rule, the models instead appear to draw on broader contextual and distributional patterns to infer syntactic categories. In sum, the results support the idea that children might be able to form robust grammatical representations even from sparse or constrained input, and underscore the value of computational modeling in investigating the mechanisms that underlie early language acquisition.

## Introduction

Learning to use grammatical rules is essential when acquiring a language. There is an ongoing debate in the field of first language acquisition about whether children have representations of these abstract rules from birth. For many years, researchers have argued that even very young children have at least some abstract syntactic knowledge. This generativist view, first introduced by Chomsky (1965), states that universal principles underlie the grammatical rules that make up a language.

More recently, researchers have started arguing for a new perspective that claims that children start off memorizing familiar items, after which they form generalized rules from overlapping sequences (Tomasello, 2003). This lexical constructivist view states that the underlying rules of a language are learned item-based first, which means no pre-existing grammatical categories are used. Adult-like categories are only learned with time, after sufficient linguistic experience (Pine et al., 2013).

While these two views seem to oppose each other, empirically showing when or if children start productively using language—thus making use of their syntactic categories—has proven to be difficult. After all, when a child produces a 'new' utterance – for example, 'the cat' when it only heard 'a cat' – it is near impossible to say that this sequence of words was not memorized from previous input. This problem led researchers to seek new ways to investigate the acquisition of grammatical knowledge. In particular, advances in Artificial Intelligence and computational modelling opened new possible ways of simulating children's language use in a controlled manner. These types of models allow researchers to directly assess how certain input can lead to rule-like behavior, when no prior input and no prior grammatical knowledge is present.

In this paper, I aimed to shed more light on this topic by computationally modelling the acquisition of grammatical knowledge—specifically the use of determiners—in English-learning children. The goal was to investigate whether these grammatical categories could be formed even when the child-directed input was constrained. By analyzing a model's generalizability under such conditions, I sought to draw conclusions about the type of input needed to support early grammatical development. The results showed that the models were able to generalize determiner use even with limited input,

suggesting that abstract grammatical categories can emerge from distributional patterns rather than direct exposure to varied determiner–noun combinations.

## Theoretical background

An important claim of the generativist theory is that children have adult-like abstract syntactic categories from the earliest observable stages, the Universal Grammar (UG) (Chomsky, 1965). According to this view, the UG provides a set of structural principles that underlie every human language. Therefore, grammatical rules are not learned from linguistic input; the learned language simply triggers a set of rules that was already present. Generativist researchers argue that even children's early utterances make use of syntactic categories (Valian, 1986; Pinker, 1996). Others argue that children, while having the correct grammars, do not use adult-like language because of performance reasons (Bloom, 1990). A detailed overview of the main arguments in favor of this view is presented in a paper by Dąbrowska (2015), including claims that all human languages share properties and thus rely on universals from birth, that children get different input but settle on the same grammar and that children acquire knowledge for which they receive no direct evidence. Dąbrowska, however, argues that these arguments are based on false or unsubstantiated premises.

In contrast, lexical constructivists argue that children's early grammatical knowledge comes from linguistic experience, and grows after more exposure to language. Rather than assuming innate knowledge, which some researchers argue is limited at best (McClure et al., 2006), this perspective holds that children first learn item-based representations, meaning that they memorize specific combinations of words that occur together frequently (e.g. "the cat"). After a while they begin to abstract over the patterns of the language they are exposed to, extracting general rules – for instance, learning that "the" is part of a determiner class and "cat" a noun class.

From this perspective, the apparent early productivity of children may thus actually be memorized item pairs masked as grammatical knowledge (Tomasello, 2003). Tomasello exemplifies this with 'verb island' constructions: children do not learn an abstract rule like Subject – Verb – Object (SVO), instead remembering specific, concrete constructions for one verb. Tomasello considers the verbs *draw* and *hit*: a child might memorize verb-specific frames like '[someone] draw' and '[someone]

hit [something]' instead of a general structure. While being able to seemingly produce 'productive' grammatical sentences using these constructions, the structure that is used for *hit* is not a generalized one, since it is not able to use that construction for *draw*. The child therefore has remembered verb-specific frames for *hit* and *draw*, instead of creating a general rule. Thus, young children are able to use determiner-noun pairs (Tomasello, 2003; Valian, 1986; Valian et al., 2009; Pine & Lieven, 1997; Alhama et al., 2025), but they don't seem to have knowledge about the relationship between different determiner types, and thus do not have a syntactic category for determiners (Pine & Lieven, 1997; Pine et al., 2013). Only after children have been exposed to a lot of linguistic input, they start to develop actual rule-based categories for those function words (Kemp et al., 2005). The use of function words with novel nouns then gradually increases as the child grows older and has even more exposure to language. This may mean that input frequency and distributional patterns are important factors in shaping language development.

An informative test case in the debate between generativists and lexical constructivists is the acquisition of English determiner-noun combinations. These constructions are one of the earliest function words children start using in the English language (Valian, 1986), they carry relatively little semantic value, yet they provide a good framework for examining children's syntactic category learning. One important characteristic is that determiner-noun usage is highly consistent, and normally has no violations of its regular pattern. Testing children for the presence of this syntactic category is often done using some measure of generalization: if a child hears 'the dog', and later is able to produce 'a dog', this suggests that they have a determiner category (Pine et al., 2013; Yang, 2011). In contrast, if such generalizations do not occur, it may indicate that children relied on memorizing word pairs instead of applying a rule. Therefore, this *productive* determiner-noun usage can offer evidence to the debate.

However, researchers disagree on what productive use reveals about any underlying syntactic knowledge, as well as how to define productivity. Pine et al. (2013) use the *overlap score* to demonstrate that children initially show limited overlap in the nouns they use with both 'a' and 'the', after which this usage increases gradually over time. From this, they infer that categories like 'determiner' are at first learned item-based from the input, meaning no (or limited) innate knowledge has to be present. Yang

(2011), however, uses the *Tolerance principle*, a mathematically formal model, to show when children have acquired a category for determiners. He demonstrated that children pass that threshold quite early, even with limited input. He also showed that this finding is unique to humans, by comparing his results with Nim Chimpsky, a chimp who learned sign language. Therefore, he argues that linguistic input *triggers* syntactic categories that were already present, instead of creating them. Thus, while Pine et al. see the acquisition of syntactic categories as an experience-based process, Yang argues that children already possess these categories early on.

Many of these studies on determiners and nouns use corpus data of child-directed or child-produced speech (Valian et al., 2009; Pine et al., 2013; Yang, 2011) and use the child-directed input and child-produced output, sometimes alongside other evidence, to determine productivity. For example, if the child-directed speech only contained 'a dog', but the child produced 'the dog', that can be seen as productive usage. An important caveat with such methods is that it is impossible to know for sure if a child has never heard a specific determiner-noun combination before the conversations in the corpus were being recorded. Ruling out memorization can therefore be quite difficult. One promising approach to address this issue is testing children's acquisition of grammatical categories through computational modelling. By simulating children's language learning with closely controlled input, parallels can be drawn to language acquisition in the real world. For example, Abu-Akel et al. (2004) used a growth modelling approach to investigate individual variation in the onset of article use in young children and McCauley & Christiansen (2019) used a Chunk Based Learner to model children's use of lexical frames: patterns of words that co-occur frequently and can provide a structure for grammar learning. Such studies illustrate how computational methods can help uncover how young children develop linguistic mechanisms.

Early computational models made for language acquisition often relied on Recurrent Neural Networks (RNNs), which are designed to process sequential data. Pioneered by Elman (1991), these types of models were able to capture aspects of syntactic dependencies implicitly, meaning they are able to learn how words and phrases relate to each other in a sentence. This is important in language modeling, since language depends quite heavily on word order and hierarchical structures. However,

these RNNs have limitations. Most notably, they struggle with long-range dependencies: words or phrases that are far apart but are syntactically or semantically related. In the past few years, Transformer-based models have therefore all but replaced RNNs in computational research of language acquisition. The specific architecture of a Transformer model allows it, crucially, to capture those long-range dependencies between variables (Vaswani et al., 2017). Vaswani et al. further showed that Transformer models outperform Recurrent and Convolutional Neural Networks, while also being more parallelizable and requiring less training time. These characteristics make Transformer models very useful for modelling natural language, since it can learn more complex patterns from less input data.

One specific instance of a Transformer model suitable for language modeling is BERT (Devlin et al., 2019). BERT, or Bidirectional Encoder Representations from Transformers, is a model structure that considers context left and right of a word simultaneously, instead of only left-to-right or right-to-left. This lets it capture more syntactic relationships than other types of model. During training, BERT often uses Masked Language Modeling (MLM): it randomly masks some words from the input, and learns to predict these masked words. This setup encourages building internal representations of the structures that underlie language. Furthermore, the model uses self-attention mechanisms, which allow the model to find the importance of every word in a sentence when it is processing a given word. This property is what allows for finding the long-range dependencies. Another important feature is that the BERT models can be trained from scratch, meaning it is possible to train a model with no inherent biases and no prior knowledge.

In recent literature, Alhama et al. (2023) trained such a BERT model from scratch on corpora of transcribed audio-recordings between children and their primary caregivers, using an MLM objective. The model was then tested by feeding it utterances produced by the children, where the determiner was masked. They found that the model started ‘productively’ using determiners at approximately the same time as children did. The model also used determiner-noun combinations it had not seen during training, meaning that it had created some generalizations from the input data. Therefore, Alhama et al. argue that these results suggest that it is possible for children to become linguistically productive from only linguistic input, and without the use of pre-existing syntactic categories.

In a follow-up study, Alhama et al. (2025) shed light on the exact onset of linguistic productivity; when does a child go beyond its input, creating novel utterances? Now only focusing on the onset measure (using at least two different nouns with both 'a' and 'the'), they used the same BERT model to find more support for the claim that children are creatively producing language beyond their input, as well as being able to identify an average age at which children begin productively producing determiner-noun combinations. This study focused on individual developmental timelines, showing that the model can accurately predict not only aggregate patterns but also individual trajectories.

However, in the discussion of this paper, Alhama et al. note the absence of information about the type of input computational models need to make accurate predictions: "[…] we do not know how little input, and what type of input, are necessary for the model to predict these productive forms" (p. 7). This observation highlights a limitation in current computational studies of language acquisition: even though certain models can create output that resembles that of a learning child, it is unclear what aspects of the input it receives allow for the generalizations it makes.

Numerous other studies point out the importance of the type of child-directed input for language learning. For example, Redington et al. (1998) showed that distributional information – information about the contexts in which a word occurs - can be extracted and employed by children to create syntactic categories. Rowe (2012) demonstrated that the quality and quantity of caregiver input significantly predicts children's vocabulary development. These studies suggest that language learning hinges on statistical patterns in the child-directed input, rather than relying on only innate knowledge of grammar. Which parts of the input that children receive enables this, is currently not entirely clear.

In the present study, I aim to fill this gap by investigating what input is necessary to generalize grammatical categories, specifically determiners, from linguistic input. The same data that Alhama et al. (2023) used to train their model will be used here (LDP corpus, see Goldin-Meadow et al., 2014), as well as roughly the same model. However, the child-directed input is modified such that every noun only appears with either type of determiner ('a boat' and never 'the boat' or vice versa). This model will be compared against a model trained on data with no restrictions, as well as three 'intermediate' models where nouns that originally were seen with both determiners in the data, were given a random chance

to be seen with only one during training. This study addresses the following question: Can a Transformer model generalize determiner use from constrained input data?

By removing variability from the training data, it will be increasingly harder or even impossible for the resulting models to rely on co-occurrence patterns between determiners and nouns to create general rules for determiner usage. Generalization here is thus defined as the creation of new determiner-noun pairings – if the noun 'boat' only occurs with 'a' during training, any use of 'the boat' during testing reflects a generalization. If a resulting model creates new determiner-noun pairings, this then has to be the result of the model generalizing syntactic categories from the input, not from memorization.

I hypothesize that models trained on more restricted data will have a lower level of generalization than the less restricted models, but will still show some. Without being able to see examples of both determiners with one noun, it seems unlikely that a restricted model would be able to learn the determiner category as well as an unrestricted model would. However, since children (and therefore probably models) can use distributional and context information to form grammatical categories (Redington et al., 1998), the co-occurrence of determiners with nouns may not be all that is needed for extracting syntactic information. With this study, I aim to shed light on the question of whether syntactic category forming can still occur when the input data is constrained. I will also discuss what this might reveal about the mechanisms children use to acquire these categories.

## Methods

### Data

This study utilizes two datasets. The primary dataset is the Language Development Project (LDP) corpus (Goldin-Meadow et al., 2014). This corpus consists of conversations between 64 children and their primary caregivers, which were videorecorded and transcribed for a maximum of twelve 90-minute sessions. The videos were recorded at the children's homes, in natural spontaneous situations. The participants in this corpus are typically developing, English speaking children from Chicago, as well as their primary caregivers. Ages of the children in this corpus ranged from 14 to 58 months. The conversations in the videos resulted in a corpus of over one million transcribed utterances ($n$ = 646,685

for the primary caregivers, $n$ = 368,884 for the children). For privacy reasons, all words that could lead to a person being identified—predominantly names—were replaced by the word 'MASK' before the data was preprocessed.

The second dataset, used only during preprocessing, is a dataset from Kaggle, consisting of popular children's names in the United States from 1880 to 2020 (Burnsworth, 2023). This dataset contains only aggregate statistical information and no personal data.

**Ethical considerations**

This study received ethical approval from the Ethics Committee of the Faculty of Humanities at the University of Amsterdam (Universiteit van Amsterdam). The study did not involve any direct interaction with human participants, since it involved only analyses on existing data. Access to the LDP corpus was granted with explicit permission from the original research team (Goldin-Meadow et al., 2014). The original team was responsible for all primary data collection procedures, including obtaining informed consent from caregivers for both their own and their children's participation, in accordance with their institutional ethical guidelines. The first names dataset is licensed under the Open Database License (ODbL, v1.0). This research is purely computational, and poses no risk to the original participants. All data was stored securely and used solely for the purposes of this academic research. In accordance with data management of personal information, the LDP corpus data was deleted upon completion of the project.

**Preprocessing**

First, all utterances had their extraneous punctuation and capitalization removed and were lemmatized. Second, each word was part-of-speech tagged using the spaCy library, specifically the *en_core_web_sm* model (Honnibal et al., 2020). Words holding personal information were previously replaced by a MASK token, which could influence model learning. Most of the MASK tokens seem to have been first names; either of the children, relatives or others. Therefore these words were replaced by a random name, taken from a dataset of the most popular first names in the United States (Burnsworth, 2023). This ensured little semantic meaning got lost in anonymizing the data. Utterances

that consisted of only MASK tokens were removed completely. Word constructions with a '+' in between were split up into two words (e.g. 'thank+you' to 'thank' and 'you'). In some utterances a combination of symbols was used to show that a word was repeated a number of times; this symbol was replaced with the repeated word, for the indicated amount of times. Table 1 shows the number of training sentences per model and per sub-dataset. All scripts used for preprocessing can be found online.

**TABLE 1 - TRAINING SENTENCES PER CONDITION**

| ***Age in months*** | **Unrestric-ted model** | **75% model** | **50% model** | **25% model** | **Restricted model** |
|---|---|---|---|---|---|
| *14* | 60394 | 60244 | 59971 | 59609 | 59202 |
| *18* | 61574 | 61416 | 61119 | 60768 | 60306 |
| *22* | 57478 | 57325 | 57037 | 56589 | 56151 |
| *26* | 55727 | 55562 | 55302 | 54896 | 54357 |
| *30* | 54021 | 53793 | 53581 | 53063 | 52541 |
| *34* | 49842 | 49685 | 49418 | 48863 | 48332 |
| *38* | 55486 | 55305 | 54970 | 54481 | 53935 |
| *42* | 50462 | 50257 | 49982 | 49536 | 48976 |
| *46* | 44565 | 44362 | 44121 | 43791 | 43294 |
| *50* | 43894 | 43726 | 43417 | 43110 | 42566 |
| *54* | 36752 | 36571 | 36340 | 35975 | 35566 |
| *58* | 36784 | 36608 | 36393 | 36057 | 35586 |

## Computational model

For the goals of this research, a model was needed that can accurately simulate children's syntactic category learning. Alhama et al. (2023) found that specifically a BERT model can closely resemble this type of learning when trained on child-directed input. They also point out that when choosing a modeling approach, there are a few important considerations. There must be no prior syntactic categories present in the model before training, similar to how constructivist theories view the earliest language learning (Tomasello, 2003). Second, the model must only have access to input data that children also have access to. Third, the task the model will be trained for must be similar to children's language experience in daily life. Alhama et al. (2023) discuss these considerations and explain that a BERT model trained from scratch with an MLM objective on the LDP corpus data fits these criteria well. Therefore I closely followed their modeling setup.

The models used in this study are small Transformer-based models, specifically BERT models (Devlin et al., 2019). I used the HuggingFace library to initialize the models (Wolf et al., 2020). The number of attention heads is set to 2 and only a single hidden layer is used. The rest of the model parameters were kept at HuggingFace's default values. These relatively limited parameters were chosen as to not overfit to the training data, since there is quite little data available, especially for the models simulating younger ages of children (modern BERT models are often trained on billions of words). The training input was tokenized using a custom BertWordPieceTokenizer (Huggingface, 2020) with a maximum vocabulary size of 30000 and a minimum frequency of 2. This tokenizer was trained from scratch on all of the caregiver's utterances. For the training parameters the batch size was increased from 8 to 64, the number of epochs was set to 4 to get initial prediction results closer to those of Alhama et al. (2023), *save_steps* was set to 10000, *save_total_limit* set to 2, *prediction_loss_only* set to TRUE, *resume_from_checkpoint* None and some logging arguments were added to track the loss. Furthermore, the Adam optimizer was used. As stated before, no syntactic categories must be present before training, thus no pre-training was applied on this model.

**Experimental design**

Multiple datasets were created from the child-directed utterances in the data. These datasets can be divided into five conditions: one condition contains the 'natural' data where no manipulations on the data were performed: the unrestricted condition. Another condition had the data manipulated such that each noun was paired with one type of determiner, making sure that across these datasets no sentences were included where a determiner of the 'wrong' sort was present: the restricted condition. This was done by first checking which determiners a noun was seen with at all in the data, then randomly 'assigning' one of these to the noun. There are also three intermediate conditions, where each noun that was seen with both determiners in the original data, has a chance of being assigned both (thus still being allowed to be seen with both determiners in the training data, instead of with one). This chance of being assigned both determiners was 25%, 50% and 75% for the intermediate conditions (called the 25, 50 and 75 condition respectfully).

After assigning the determiners to the nouns, every determiner-x-noun pairing (where x was 0 – 2 words other than a determiner or noun) was checked to see if it was consistent with the assigned determiner(s). If not, the determiner was changed to the correct type. Some words that were seen at one point as a noun, and thus had a determiner assigned to them, could be seen as another part of speech somewhere else. An example is "I would like a large" and "the large mountain", where 'large' is a noun in the first sentence and an adjective in the second. If large was assigned the determiner 'a', the second sentence has 'the' in front of 'large' and thus the determiner still was changed, even when 'large' is not a noun in that case. If two words in a sentence like "the large mountain" wanted conflicting determiners ('large' is assigned to 'a' and 'mountain' is assigned to 'the'), the sentence was removed from the training data for simplicity.

Next, 12 datasets were created per condition. The first dataset contained all the child-directed utterances up until the age of 14 months, the second used the data after that until the maximum age to 18 months, the third to 22 months, etcetera. This resulted in 5 types of datasets, each with 12 sub-datasets per age range. These sets will serve as the training data for the models. Training set sizes can be found in Table 1.

The test sets consisted of utterances produced by the children. These were first filtered such that only utterances remained where a determiner-noun or determiner-x-noun pairing was used, where x is any word other than a determiner or noun. Then, the utterances were filtered such that only sentences remained where a noun was used that was also present in the training data, where it was seen with just 1 type of determiner. This means removing sentences containing a determiner-(x-)noun where the noun was either not seen in the training data for that model, or seen with both determiners in the training data. This is to ensure that all utterances in the test data contain a noun that has been seen before by the model, while also giving the model the opportunity to create a pairing where the before-seen noun can be combined with a determiner it was not paired with before. Every noun was only allowed to be seen once per test set, as to not have the results be obscured by nouns that had much higher frequencies. Simply testing the models on bare determiner-noun constructions resulted in the models overpredicting the 'a' determiner, hence the current setup.

All models were trained with an MLM objective. 15% Of the words in each input utterance were masked using the DataCollatorForLanguageModeling from the transformers library (Wolf et al., 2020). Training was done incrementally on subsets of caregivers' utterances from the LDP corpus. For the subsets, utterances for children were bundled since individual children's data is not enough to train a model. Each subset contained all utterances directed to children, incrementing from the previous age limit up to the current age limit (starting with data of up to 14 months and increasing in 4-month intervals, ending at 58 months). Incrementally training the models therefore mirrors the real-world accumulation of linguistic experience of children.

**Evaluation**

Model evaluation was mostly focused on the model's ability to generalize grammatical rules about determiner use from the input data. The test sets consisted of all utterances that were produced by children, and contained at least one determiner-noun pairing, of which the noun was present in the training data with only one determiner. This allows for direct testing of whether the model is able to generate novel determiner-noun pairings, without being able to rely on rote memory.

During testing, every determiner was masked, and the model was tasked with predicting the masked word. For example, the sentence *You give me a toy* would be changed to *You give me [MASK] toy*, with the model filling in the masked slot. A model's ability to generalize was evaluated based on the amount of unseen determiner-noun combinations it made during testing. If the model predicted 'a' when it only saw the paired noun with 'the' in the data (or vice versa), the model showed a generalized use of that determiner.

As a means of measuring the models' productive determiner use and comparing it against that of Alhama et al.'s (2023), the productivity onset of the models was determined. For this, a new test set was created. This set contained all children's utterances that contained a single determiner-noun or determiner-x-noun construction, with each utterance having a label for the age in months and the subject number. Models had to predict the masked determiner, after which the onset measure was calculated per individual child and age. Onset measure is defined by Alhama et al. (2023) as a child or model using

both ‘a’ and ‘the’ for a single noun, twice. This is a useful measure to indicate when children go beyond using rote memory, thus using grammatical rules for producing language.


**FIGURE 1**
UNRESTRICTED MODEL TRAINING LOSS.

## Overall performance

As can be seen in the loss plot of Figure 1, the unrestricted model learned well and converged after 4 epochs. This trend is very similar for all models (see Appendix A). The increases in loss are the result of the loss ‘resetting’ when a model was saved, then loaded and further trained on the next sub-dataset. To assess if the unrestricted model’s training was successful, it was tested on all children’s utterances that contained a determiner-noun or determiner-x-noun construction, where the determiner was masked and predicted. Figure 2 shows a confusion matrix of the fully trained unrestricted model (A), as well as the confusion matrix from Alhama et al. (2023) (B). The model trained in this study has very similar prediction patterns, although with a slightly lower overall accuracy and determiner prediction rate (61% versus 65% and 79% versus 83%, respectively).

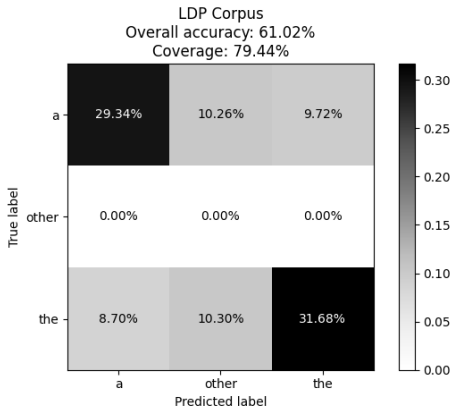



**FIGURE 2A**
CONFUSION MATRIX OF THE UNRESTRICTED MODEL PERFORMANCE.

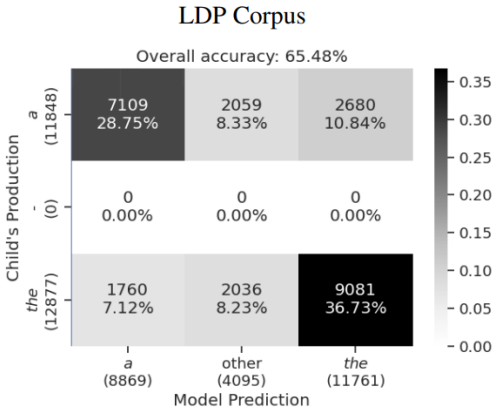



**FIGURE 2B**
FIGURE FROM ALHAMA ET AL. (2023); REPRODUCED HERE WITH PERMISSION FROM THE AUTHORS.

## Results

In this section I go over the specifics of the datasets and evaluate performance of the models on the curated test sets, focusing on their ability to generalize determiner use, prediction accuracy and onset of productive determiner use.

### Data characteristics

Across the original dataset, there were 5507 unique nouns, of which 1506 were seen with only determiner ‘a’, 1933 with only ‘the’ and 2068 with both ‘a’ and ‘the’. For every condition, the average amount of nouns appearing with both determiners per sub-dataset was calculated. Table 2 shows this metric.

**TABLE 2 - AVERAGE NUMBER OF NOUNS APPEARING WITH TWO DETERMINERS IN SUB-DATASETS**

| ***Model type*** | **Average number of nouns with two determiners (% of unrestricted)** |
|---|---|
| *Unrestricted* | 431 (100%) |
| *75 condition* | 341 (79%) |
| *50 condition* | 247 (57%) |
| *25 condition* | 122 (28%) |
| *Restricted* | 0 (0%) |

**Generalization of Determiner Use**

All models were able to predict ‘a’ for nouns that were seen with only ‘the’ in training and vice versa. Figure 3A demonstrates that the rates at which they do this is much above zero. This pattern even held for the models trained only on the data of young children, as well as for all different conditions. Generalization seems to increase with models trained on more increments. However, this trend can be ascribed to the test sets becoming bigger for later models, since those models had seen more nouns, allowing more nouns to be suitable for the test set. The models trained on more restricted data had more testing data as well, since in those conditions there were more nouns that were seen with only one determiner in the training data. This might explain the higher total amount of generalizations in the more restricted models.

To address this test size issue, generalization ratios were calculated. Figure 3B shows the ratio of new determiner-noun pairs made by the model, to the total amount of determiners it predicted. Specifically, it shows how often the model predicted determiner ‘a’ when it only saw ‘the’ in training for the noun in the test sentence (and vice versa), compared to how often it predicted a determiner at all. This metric is also shown in Table 3. Even when controlling for test size, all models are able to generalize determiner use. And again, generalized determiner use is much above zero for every type and ‘age’ of model. Table 3 shows that the average generalization rate for the most restricted model seems slightly lower than the other types of model at 32% . However, Figure 3B demonstrates that all models end up with a 30-40% generalized prediction ratio when fully trained.

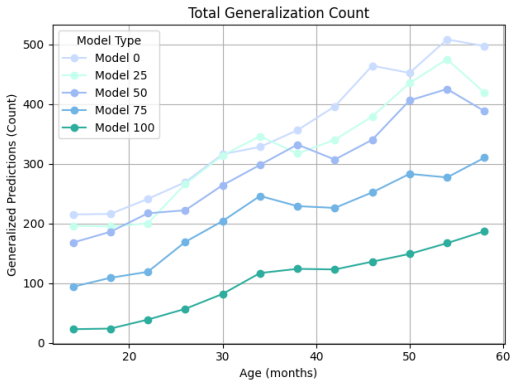



**FIGURE 3A**
TOTAL OF GENERALIZED PREDICTIONS PER MODEL. MODEL 0 REFERS TO THE RESTRICTED MODEL, MODEL 100 REFERS TO THE UNRESTRICTED MODEL.

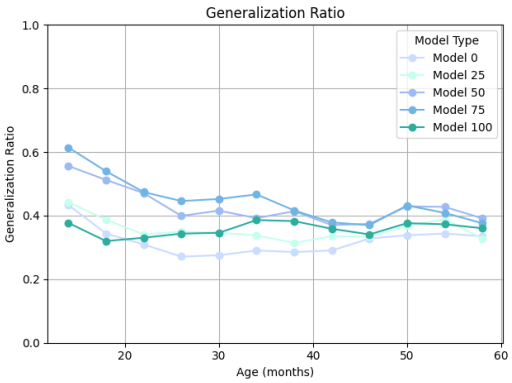



**FIGURE 3B**
RATIO OF GENERALIZED PREDICTIONS TO TOTAL DETERMINER PREDICTIONS PER MODEL.

**Determiner predictions and accuracy**

Figure 1 showed that all models have learned during training. However, to investigate the possibility of the above results being a consequence of the models not having learned the correct features, the prediction ratios of the determiner class was computed. All models were very similarly able to determine when to use a determiner, and used both ‘a’ and ‘the’ at approximately the same rate (see Figure 4A). This determiner prediction ratio was above chance level (50%) for all models, and at all ages. The figure also shows that all types of models follow the same general pattern in determiner class prediction very closely. They start off around 60 percent, do not learn much in the first 2 increments and then have a sudden ‘learning spike’, after which the trend reaches a plateau.

Model accuracy was also evaluated, the results of which are shown in Table 3. Overall accuracy refers to the proportion of total predictions in which the model correctly identified the exact masked token. Accuracy on determiner prediction (conditional accuracy) is the proportion of predictions of the exact masked token, given that the model predicted any determiner. This last metric is also illustrated in Figure 4B. Both the overall accuracy and conditional accuracy were lower the more restricted the model’s training data was. For example, the model where no noun was seen with both determiners did not surpass 60% conditional accuracy (after the first increment), while the regular model—trained on

unmodified data—reached conditional accuracies closer to 80%. Excluding the most restricted condition, accuracy was much above chance level (50%) for all models and ages.

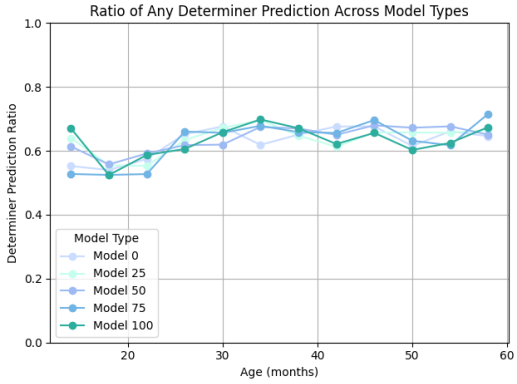



**FIGURE 4A**
RATIO OF ANY DETERMINER PREDICTIONS TO TOTAL PREDICTIONS PER MODEL.

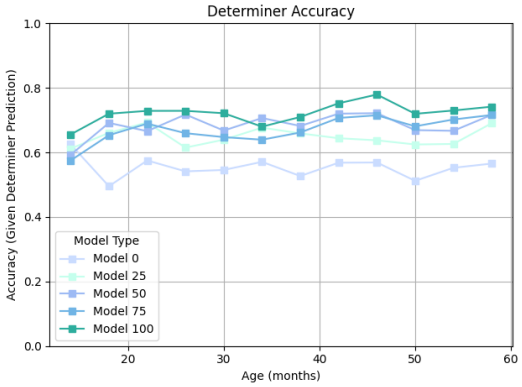



**FIGURE 4B**
RATIO OF CORRECT DETERMINER PREDICTIONS TO TOTAL DETERMINER PREDICTIONS PER MODEL.

**TABLE 3 – SUMMARY STATISTICS PER MODEL TYPE**

| *Model type* | **Mean generalization ratio** | **Sd generalization ratio** | **Accuracy on DET prediction** | **Sd on DET prediction** | **Mean overall accuracy** | **SD overall accuracy** |
|---|---|---|---|---|---|---|
| *Restricted* | 0.32 | 0.045 | 0.55 | 0.034 | 0.35 | 0.033 |
| *25%* | 0.36 | 0.035 | 0.65 | 0.028 | 0.41 | 0.031 |
| *50%* | 0.43 | 0.057 | 0.68 | 0.037 | 0.44 | 0.040 |
| *75%* | 0.45 | 0.071 | 0.67 | 0.040 | 0.42 | 0.060 |
| *Unrestricted* | 0.36 | 0.022 | 0.72 | 0.032 | 0.46 | 0.036 |

GENERALIZATION RATIO IS CALCULATED AS THE RATIO OF GENERALIZED PREDICTIONS TO TOTAL DETERMINER PREDICTIONS. ACCURACY ON DET PREDICTION IS THE CONDITIONAL ACCURACY; THE PROPORTION OF CORRECT PREDICTIONS, GIVEN THAT THE MODEL PREDICTED ANY DETERMINER. OVERALL ACCURACY IS THE ACCURACY ACROSS ALL PREDICTIONS.

## Productivity onset

Figure 5 shows the median number of times the unrestricted model predicted both ‘a’ and ‘the’ for a single noun, as well as this same measure for the individual children in the dataset. The model shows the same general pattern of productive use of determiners as the children, although using them productively slightly more from the age of 26 months onward. These patterns are similar to those found in the paper of Alhama et al. (2023) (see Figure 6), however, the current model is more productive in determiner use than Alhama’s model.

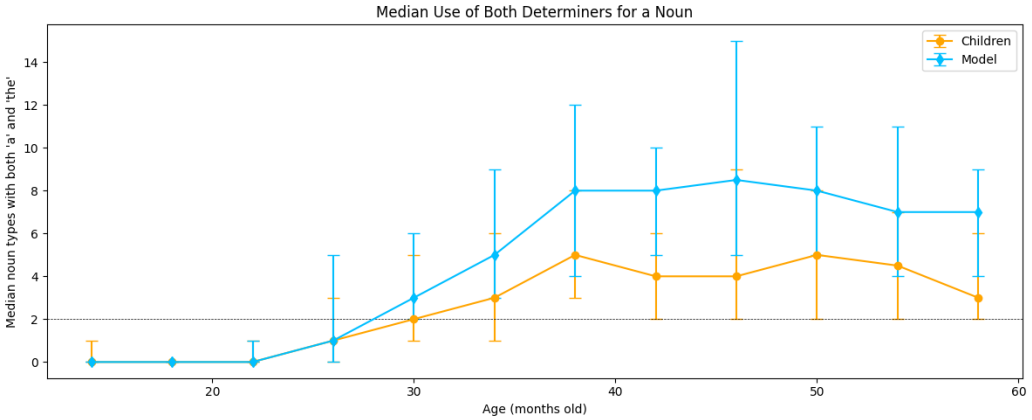



**FIGURE 5**

MEDIAN NUMBER OF NOUN TYPES THAT WERE USED WITH BOTH DETERMINERS. HORIZONTAL LINE REPRESENTS ONSET CRITERION USED IN ALHAMA ET AL. (2023). ERROR BARS CORRESPOND TO STANDARD DEVIATION.

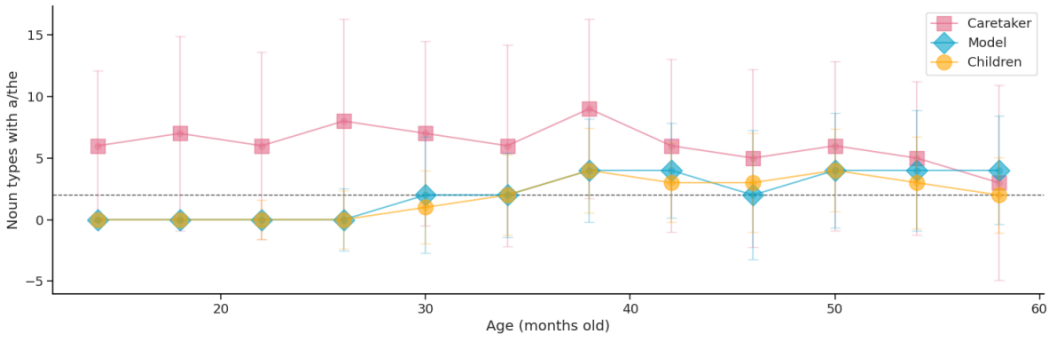



**FIGURE 6**

MEDIAN NUMBER OF NOUN TYPES THAT WERE USED WITH BOTH DETERMINERS. FIGURE FROM ALHAMA ET AL. (2023); REPRODUCED HERE WITH PERMISSION FROM THE AUTHORS.

## Discussion

To examine how restricting the variation of determiner–noun pairings in training data influences a model's ability to generalize determiners, this study used a Transformer architecture, which was previously shown to accurately reproduce children's productivity in determiner-noun usage (Alhama et al., 2023). Building on the modeling approach of Alhama et al., this study introduced specific modifications to the input data to test how limiting determiner–noun pair variability affects determiner generalization. Results showed that all models were able to generalize determiner use, and used these generalizations often.

**Generalization in all models**

A first notable observation is the fact that all models were able to use a determiner for a noun other than the one that noun was seen with in the training data. Alhama et al. (2023) already showed that an incrementally trained Transformer model on the LDP corpus was able to go beyond its input, thus creating generalized rules and applying them. However, here I show that even a model that does not have access to examples of nouns with different determiners, still is able to acquire the general rule that both determiners can be used with a noun. This suggests that generalization is not reliant on observing variability in determiner-noun combinations, but may come from other, broader, linguistic cues.

These results reflect findings in human infants. Somewhat similar learning patterns to those in this study were found in Gomez and Gerken (1999), where 12-month-old infants were exposed to a small subset of sentences from an artificial language. Training sentences were generated using a finite-state grammar, after which the children were tested by having them discriminate grammatical from ungrammatical strings, as well as discriminating a grammar's strings from those of a different but similar grammar. The children were also tested for recognition of a learned grammar when a different vocabulary (but the same grammar) was used. Even when training was brief and based on a limited vocabulary, children showed that they were able to learn rules like finite-state structures, non-adjacent dependencies and variable word order.

Moreover, Goldin-Meadow and Mylander (1998) showed that deaf children whose access to regular linguistic input was limited, still created structured gesture systems (homesign). They found that these gesture systems did not consist of just random signs, but rather showed consistent ordering of gestures, and had combinations of patterns and segmentations of sentence-like units. This syntactic-like structure emerged without the aid of consistent grammatical linguistic input. These studies underscore an important point: children seem to be able to extract syntactic knowledge from input that is sparse or limited, whether due to a small vocabulary, restricted pairings or a lack of regular language. This can happen even for children as young as 12 months old.

This raises a fundamental question: what, then, allows a child or model to abstract syntactic categories from the input it is getting? The current results would suggest that generalized rule use does not have to be learned from instances of the rule itself, since in the restricted condition the determiner-noun pairing was not a good indicator of the underlying rule (that it is possible to use both determiners for a noun). Thus, models do not seem to rely heavily on rote memory, or even specifically on examples of the rule being used for learning how to use determiners. That means contextual cues have to be employed to extract syntactic categories.

These observations seem to be in line with research findings indicating that distributional information could play a significant role in acquiring syntactic categories (Redington et al., 1998; Mintz, 2003; Reeder et al., 2013). This means that grammatical knowledge of the determiner class could be extracted from the distribution of contexts the class is found in. While Redington et al. primarily focused on sequential order information (word order), Mintz's 'frequent frames' model demonstrates that infants can use context-based cues to identify categories. The models in this study may thus have been using the context of the determiner-noun pairings to figure out the rule behind using them. Future studies should explore which specific context features support generalization in model learning.

**Early generalization**

Second, generalization occurs even when models are only trained on data of the youngest children. At first glance, this seems to be in contrast with studies that take on a lexical constructivist view, since they argue that grammatical categories are not present in young children, and are learned from linguistic experience (Tomasello, 2003; Lieven et al., 1997). Since even all 'young' models were often able to determine when to use a determiner, and generalized the use of them, this would seem to argue against the absence of syntactic categories at young ages. Importantly, however, the models did not start off with any grammatical knowledge since they were trained from scratch. The generalizations and rules they picked up had to have come from patterns in the data.

This finding might indicate that while young children may not have an innate understanding of syntactic categories, they might be able to use distributional patterns and learning mechanisms to extract grammatical knowledge from linguistic input from very early ages, long before being able to use the

knowledge productively. This would be in line with findings from Lany and Saffran (2010), who found that infants are able to learn lexical categories before they are able to speak, as well as before they have semantic understanding of words. Alhama et al. (2023) and Figure 5 show that the children in the LDP corpus and the models in this study were quite unable to use both determiners with a single noun until 30 months of age. This, in combination with the fact that all models were able to immediately generalize determiner use, seems to argue in favor of Lany and Saffran's point; the models acquired syntactic knowledge before they were able to use it. Furthermore, Saffran et al. (1996) showed that infants as young as 8 months old were able to use statistical learning mechanisms to extract linguistic information. The findings from these studies seem to be consistent with the findings in this study, where every type of model, at every age, was able to extract syntactic categories, but was not consistently able to apply the rule it learned until a later 'age'. It thus seems that more research is needed on the apparent gap between acquiring syntactic knowledge and using it, to fully understand when children obtain knowledge of grammatical categories.

**Accuracy differences**

Next, the only major difference in model behavior between the five conditions is the accuracy at which they could predict the masked slot. This means that restricting the variability in determiners per noun makes it so that models do not learn which exact determiner to use in what context. As shown in Figure 4, all models were equally able to ascertain when to use a determiner, but differed in their accuracy of predicting the exact token (determiner) that was masked. This may tie back into the studies pointing at distributional information being essential in forming syntactic categories (Redington et al., 1998; Mintz, 2003; Reeder et al., 2013). For all types of models, the context of the pairing was the same since only the determiners were changed in processing. Therefore, all models were equally able to use the context cues and distributional information to learn when to use the determiner category. However, forming a specific 'a' or 'the' category was much harder for the more restricted models, since examples of 'a' usage may be present in 'the' sentences, and vice versa. That is to say, the context of the sentence always stayed the same, but the determiner may be changed, such that the determiner now does not fit

with the context of the sentence anymore, but the model will still try to use that context to learn which determiner to use. This may explain the notable difference in accuracy for specific determiner use.

**Comparison to previous work**

Lastly, it seems that it is not possible to ascribe these results to aspects of the model architecture or training. The unrestricted model was compared against the model in Alhama et al.'s (2023) study with a confusion matrix, accuracy score and a productivity onset measure. These showed that the model in this study has very similar patterns and behaviors as Alhama et al.'s. Interestingly, however, the current (unrestricted) model seems to outperform Alhama et al.'s model and the children from the LDP corpus in productive determiner use (see Figure 5 and Figure 6). While mostly following the same general pattern, the model's median productive determiner use surpasses that of children's past 26 months. At first a plausible reason for this seems to be that the models in this study were trained with 4 epochs instead of 3. This could have given the model more time to pick up on the pattern that multiple determiners can be used for a single noun, possibly even overfitting to this pattern slightly. However, even when trained with 3 and 5 epochs, the pattern of overproducing was still similarly present (see Appendix B).

The only other difference between the two models is the small discrepancies in the way data is preprocessed. In the current study, constructions indicated with a '+' were split up into two words. Words that were followed by a symbol to represent that the word was repeated (represented with for example '[x2]') had their symbol removed, and the repeated word added. MASK words that indicated personal information were replaced with a random first name from an external database. These manipulations, to my knowledge, were not done in Alhama et al.'s setup. While unlikely that this influenced model learning that much, it is possible that the models picked up slightly different patterns since their training data was different. Due to time constraints it was not possible to better determine why the models performed slightly differently. However, even with these differences the models nevertheless performed comparably in predicting determiners.

**Limitations**

Despite the strengths of this study, several limitations should be noted. First, Transformer models are trained on linguistic input only, whereas human language learning likely uses multiple modalities and social cues, as even without regular linguistic input humans can create syntactic categories (Goldin-Meadow & Mylander, 1998). Thus, the usage of only text-based input is inherently a limitation. Second, determiners are only a part of syntactic development as a whole. While useful as a case study, findings from determiner-noun usage may not generalize to other grammatical categories or constructions as well. Moreover, many languages use determiners differently than the English language does, or do not even use them at all. The current results can therefore not be taken as evidence for all languages. Third, the manipulation of determiner variability led to somewhat unnatural input; children are not very likely to encounter linguistic input where nouns never occur with both determiners, meaning that the input for the models used is not exactly regular language. Additionally, automatic filtering of the determiner-noun pairs may have missed some instances of 'wrong' determines being present with a noun, allowing the occasional exposure to some variability even in the most restricted conditions. Lastly, although the models were tested on child-produced sentences, their behavior still is that of a computational model trained on linguistic data, and the output produced by such a model cannot be taken as direct evidence of a child's linguistic mechanisms. These limitations point to the importance of using studies on model behavior alongside child experimental studies to accurately assess children's language learning in future studies.

## Conclusion

This study investigated the impact of restricting determiner variability in training data on Transformer models' ability to generalize determiner use. This was done by replicating the setup of Alhama et al. (2023) and making some key modifications to the input of the models: per condition, the amount of nouns that were seen with both determiners in training was reduced, or even set to 0. After training, the models were tested on children's produced sentences.

The results demonstrate that the models are able to create general syntactic categories for determiners, even when they are not exposed to variable determiner-noun examples in training, suggesting that generalization is not reliant on direct exposure to the determiner-noun combinations

themselves, but can emerge from some broader distributional patterns in the input. Crucially, even the models that were trained on input directed at the youngest children were able to infer generalized determiner use, mirroring findings from other studies, that showed that children can extract grammatical categories from limited or irregular input (Gomez & Gerken, 1999; Goldin-Meadow & Mylander, 1998). This supports the idea that distributional patterns are important for creating syntactic categories. The importance of such patterns for category learning also explains the lower accuracy of more restricted models in predicting the exact determiners, since these models had more conflicting context cues to extract specific categories from. Lastly, the models seemed to behave very similarly to the models used in Alhama et al.'s (2023) study, with only the productivity rate being higher, likely because of some small changes in data processing.

Overall, these findings add to our understanding of how Transformer models acquire syntactic generalizations, and how this may relate to language acquisition in children. In both systems, syntactic categories needed for productive determiner use seem to develop before the productive use itself is possible. The fact that models could generalize determiner use from limited input supports the hypothesis that children are able to use sparse input to form syntactic categories as well, and use distributional cues to do so. This study highlights the value of computational modeling approaches for isolating input features that may be important for language learning. Ultimately, tools and studies like these bring us closer to understanding the mechanisms that enable language learners to acquire and productively use grammar from linguistic input.

## References

Abu-Akel, A., Bailey, A. L., & Thum, Y. (2004). Describing the Acquisition of Determiners in English: A Growth Modeling Approach. *Journal Of Psycholinguistic Research*, *33*(5), 407–424. https://doi.org/10.1023/b:jopr.0000039548.35396.c2

Alhama, R. G., Foushee, R., Byrne, D., Ettinger, A., Alishahi, A., & Goldin-Meadow, S. (2025). Using computational modeling to validate the onset of productive determiner–noun combinations in English-learning children. *Proceedings Of The National Academy Of Sciences*, *121*(50). https://doi.org/10.1073/pnas.2316527121

Alhama, R. G., Foushee, R., Byrne, D., Ettinger, A., Goldin-Meadow, S., & Alishahi, A. (2023). Linguistic Productivity: the Case of Determiners in English. *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*, 330–343. https://doi.org/10.18653/v1/2023.ijcnlp-main.21

Bloom, P. (1990). Subjectless Sentences in Child Language. *Linguistic Inquiry*, *21*(4), 491–504. http://www.jstor.org/stable/4178692

Burnsworth, R. (2023). *Popular names by birth year (1880-2022)*. Kaggle. https://www.kaggle.com/datasets/ryanburnsworth/popular-names-by-birth-year-1880-2022

Chomsky, N. (1965). *Aspects of a theory of syntax*. Cambridge: MIT Press.

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it?. *Frontiers in psychology*, *6*, 852. https://doi.org/10.3389/fpsyg.2015.00852

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186. Association for Computational Linguistics.

Elman, J. L. (1991). Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. *Machine Learning*, *7*(2), 195–225. https://doi.org/10.1023/A:1022699029236

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New Evidence About Language and Cognitive Development Based on a Longitudinal Study: Hypotheses for Intervention. *The American Psychologist*, *69*(6), 588–599. https://doi.org/10.1037/a0036886

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*(6664), 279–281. https://doi.org/10.1038/34646

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135. https://doi.org/10.1016/s0010-0277(99)00003-7

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Hugging Face. (2020). *Tokenizers: Fast, customizable, and extensible tokenizers*. GitHub. https://github.com/huggingface/tokenizers

Kemp, N., Lieven, E., & Tomasello, M. (2005). Young children's knowledge of the "determiner" and "adjective" categories. *Journal of speech, language, and hearing research : JSLHR*, *48*(3), 592–609. https://doi.org/10.1044/1092-4388(2005/041)

Lany, J., & Saffran, J. R. (2010). From Statistics to Meaning. *Psychological Science*, *21*(2), 284–291. https://doi.org/10.1177/0956797609358570

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal Of Child Language*, *24*(1), 187–219. https://doi.org/10.1017/s0305000996002930

McCauley, S. M., & Christiansen, M. H. (2019). Modeling Children's Early Linguistic Productivity Through the Automatic Discovery and Use of Lexically-based Frames. *Proceedings of the*

*Annual Meeting of the Cognitive Science Society*, 41. Retrieved from https://escholarship.org/uc/item/644529zb

McClure, K., Pine, J. M., & Lieven, E. V. M. (2006). Investigating the abstractness of children's early knowledge of argument structure. *Journal of Child Language*, *33*(4), 693–720. doi:10.1017/S0305000906007525

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. https://doi.org/10.1016/s0010-0277(03)00140-9

Pine, J. M., Freudenthal, D., Krajewski, G., & Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, *127*(3), 345–360. https://doi.org/10.1016/j.cognition.2013.02.006

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138. doi:10.1017/S0142716400009930

Pinker, S. (1996). Language Learnability and Language Development. In *Harvard University Press eBooks*. https://doi.org/10.4159/9780674042179

Redington, M., Chater, N., & Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science*, *22*(4), 425–469. https://doi.org/10.1207/s15516709cog2204_2

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2012). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30–54. https://doi.org/10.1016/j.cogpsych.2012.09.001

Rowe M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Tomasello, M. (2003). *Constructing a language.* Cambridge: Harvard University Press.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *22*(4), 562–579. https://doi.org/10.1037/0012-1649.22.4.562

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal Of Child Language*, *36*(4), 743–778. https://doi.org/10.1017/s0305000908009082

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). *Transformers: State-of-the-art natural language processing*. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

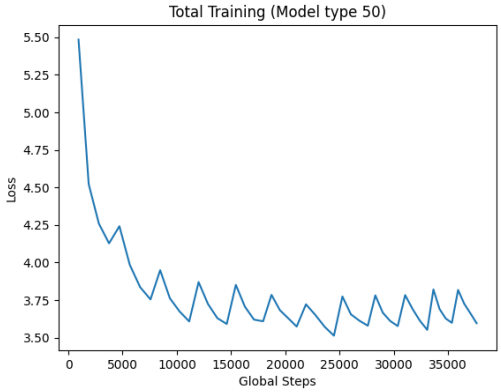
Yang, C. (2011). A Statistical Test for Grammar. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 30–38. https://www.ling.upenn.edu/~ycharles/papers/acl2011.pdf

# Appendices

## A – Loss plots per model type


**Loss plot of the restricted condition model**

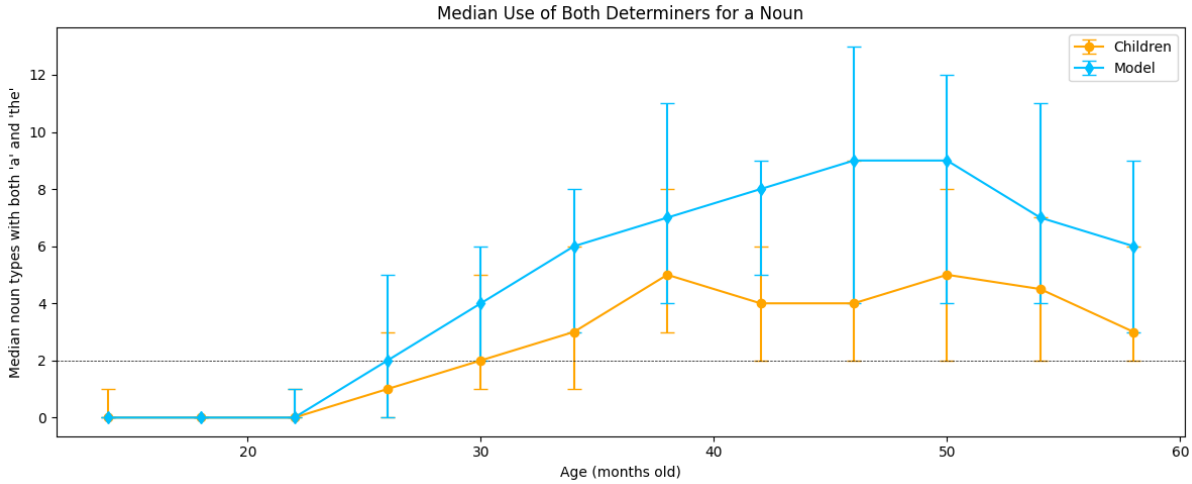

**Loss plot of the 25% condition model**



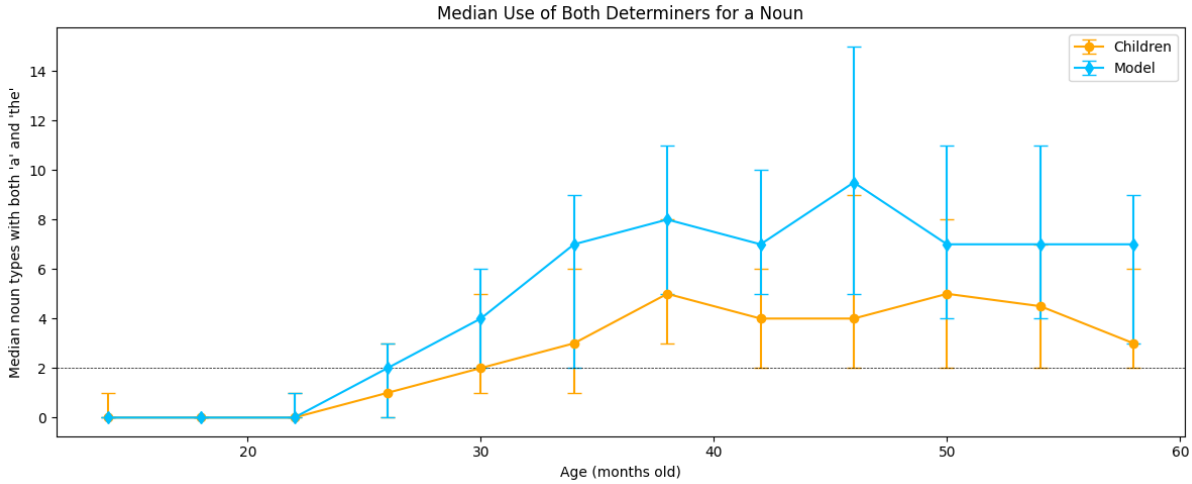

**Loss plot of the 50% condition model**


**Loss plot of the 75% condition model**

## B – Productivity onset with different epochs




**Productivity onset of unfiltered model trained with 3 epochs instead of 4.**




**Productivity onset of unfiltered model trained with 5 epochs instead of 4.**