Modeling the influence of language input statistics on children's speech production

Abstract

We trained a computational model (the Chunk Based Learner; CBL) on a longitudinal corpus of child-caregiver interactions to test whether one proposed statistical learning mechanism—backward transitional probability (BTP)—is able to predict children's speech productions with stable accuracy throughout the first few years of development. We predicted that the model less accurately generates children's speech productions as they grow older because children gradually begin to generate speech using abstracted forms rather than specific "chunks" from their speech environment. To test this idea, we trained the model on both recently encountered and cumulative speech input from a longitudinal child language corpus. We then assessed whether the model could accurately reconstruct children's speech. Controlling for utterance length and the presence of duplicate chunks, we found no evidence that the CBL becomes less accurate in its ability to reconstruct children's speech with age. Our findings suggest that BTP is an age-invariant learning mechanism.

Keywords: statistical learning, language learning, abstraction, developmental trajectory, age-invariance, CHILDES, children

Word count: 6056

Modeling the influence of language input statistics on children's speech production

During the first few years of life children learn the basic building blocks of the language(s) around them. One way they do so is via statistical learning (SL), the process of extracting regularities present in the language environment. Over the past few decades, SL has become a major topic in the field of first language acquisition, ranging in application from speech segmentation (Jusczyk & Aslin, 1995; J. R. Saffran, Aslin, & Newport, 1996) and phonotactic learning (Chambers, Onishi, & Fisher, 2003) to producing irregulars (Arnon & Clark, 2011), discovering multi-word structures (Bannard, Lieven, & Tomasello, 2009; Chang, Lieven, & Tomasello, 2006; Frost, Monaghan, & Christiansen, 2019), and much more (see J. R. Saffran and Kirkham (2018) for a recent review). By its nature, work in this domain is heavily concerned with at least two major topics: (1) the information available in children's language environments (the "input") from which they can pick up on patterns, and (2) the precise mechanisms by which children convert these "raw" environmental statistics into internalized knowledge about language. A third issue is whether and how children's SL behavior changes as they develop (Shufaniya & Arnon, 2018). The current paper taps into each of these three issues: we train a computational model on a longitudinal corpus of child-caregiver interactions to test whether one proposed SL mechanism—backward transitional probability (BTP)—is able to predict children's speech productions with stable accuracy as they get older.

SL over development

The ability to detect and store patterns in the environment begins in infancy (e.g., S. P. Johnson et al., 2009; Kidd, Junge, Spokes, Morrison, & Cutler, 2018; J. R. Saffran et al., 1996; Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009), continues into adulthood (e.g., Christopher M Conway, Bauernschmidt, Huang, & Pisoni, 2010; Frost & Monaghan, 2016; J. R. Saffran, Johnson, Aslin, & Newport, 1999), and crosses a range of modalities

(Christopher M. Conway & Christiansen, 2005; Emberson, Conway, & Christiansen, 2011; Monroy, Gerson, & Hunnius, 2017). However, it is still a matter of debate whether SL is an age-invariant skill or not (Arciuli & Simpson, 2011; Raviv & Arnon, 2018; J. R. Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Shufaniya & Arnon, 2018). Recent work that investigates SL abilities in 5–12-year-old children suggests that, while both visual and auditory SL improve with age for non-linguistic stimuli, performance stays the same across childhood for linguistic stimuli (Raviv & Arnon, 2018; Shufaniya & Arnon, 2018). From this finding, the authors conclude that SL for language might be age-invariant. On the other hand, infant SL abilities do appear to shift within the first year, both for linguistic (Kidd et al., 2018) and non-linguistic (S. P. Johnson et al., 2009) stimuli. For example, while 11-month-olds can detect and generalize over regularities in a sequence, 8-month-olds are only capable of detecting the regularities, and neither group succeeds yet at learning visual non-adjacent dependencies (see also Bulf, Johnson, & Valenza, 2011, and @slone2015infants; S. P. Johnson et al., 2009).

These changes in SL behavior during infancy and early childhood may relate to changes in other fundamental cognitive skills. For example, SL-relevant brain regions, such as the pre-frontal cortex, continue maturing through childhood (Casey, Giedd, & Thomas, 2000; Diamond, 2002; Rodríguez-Fornells, Cunillera, Mestres-Missé, & Diego-Balaguer, 2009; Uylings, 2006), which may change how children attend to the linguistic information around them as they get older. Similarly, infants' long-term memory continuously improves between ages 0;2 and 1;6 (Bauer, 2005; Wojcik, 2013). Therefore, the manner in which they store linguistic regularities in long-term memory may also shift during this period. Relatedly, working memory and speed of processing change continuously throughout early childhood (Gathercole, Pickering, Ambridge, & Wearing, 2004; Kail, 1991), implying that there could be a developmental change in the rate and scale at which children can process chunks of information from the unfolding speech signal.

Continued exposure to linguistic input itself can also be an impetus for change in SL behavior—a view supported by multiple, theoretically distinct, approaches to early syntactic learning. For example, Yang (2016) proposes that children gather detailed, exemplar-based statistical evidence until it is more cognitively efficient for them to make a categorical abstract generalization. He proposes that, at that point, the learner instantiates a rule to account for patterns in the data. Usage-based theories of early language development instead propose that children first learn small concrete linguistic sequences from their input that are made up of specific words or word combinations (e.g., "dog" and "I wanna"; or multi-word combinations, "where's the ... "; Tomasello (2008)). Then, over time, children are proposed to form abstract schemas centered on lexical items (see also Bannard et al. (2009) and Chang et al. (2006)). This representational shift, from probabilistic and lexical to abstract and syntactic, is used to account for how children can eventually create utterances that they have never heard before. Crucially, the representational shift implies a change in the way children apply the original SL mechanism(s) to incoming linguistic information (see also Lany & Gómez, 2008).

Change in SL behavior following further linguistic experience is also predicted in models that do not assume abstraction. In chunk-based models of language learning (Arnon, McCauley, & Christiansen, 2017; M. H. Christiansen & Arnon, 2017; M. H. Christiansen & Chater, 2016; Misyak, Goldstein, & Christiansen, 2012; StClair, Monaghan, & Christiansen, 2010), children use statistical dependencies in the language input (e.g., between words or syllables) to store chunks of co-occurring forms. Dependencies between the chunks themselves can also be tracked (see, e.g., Jost & Christiansen, 2016); as the outcome of continued exposure and chunk storage. Fundamentally, however, the larger chunks in these models are designed to use the same underlying mechanisms (Misyak et al., 2012), so there is no major representational shift similar to those proposed above, and the mechanisms for processing, storing, and deploying information stay the same.

We investigated the possibility of developmental change in SL using computational modeling, which enables us to define and test the goodness-of-fit for any given learning mechanism on a dataset of natural speech. We chose to use a longitudinal child language dataset, in which the same children were tracked across the developmental period of interest for early speech production (0;11–4;0). By choosing data in this age range, we could test whether use of a learning mechanism changed for each child across the studied developmental time points. We tested for developmental change in the use of a single proposed statistical learning mechanism: backward transitional probability (McCauley & Christiansen, 2011; Onnis & Thiessen, 2013; Pelucchi, Hay, & Saffran, 2009).

BTP and the Chunk-Based Learner

Our model is based on McCauley and Christiansen's (2011, 2014a) Chunk-Based Learner (CBL) model, which uses one measure—backward transitional probability—to detect statistical dependencies in the speech stream. BTP for a given pair of words is defined as the occurrence probability of the previous word (w_{-1}) given the current word (w_0). It can be estimated for each word in a sentence in order to reveal peaks and dips in transitional likelihood, which reflect places where words are likely (peaks) or unlikely (dips) to co-occur.

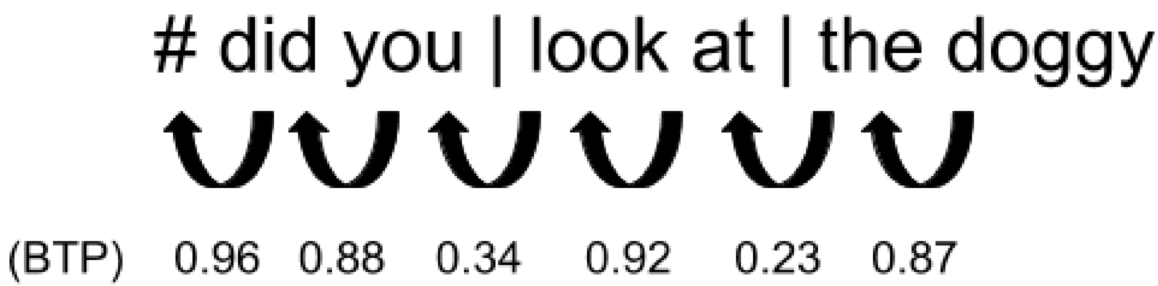

Figure 1. Example of a sentence with BTP between consecutive words. Chunks are split at points of low BTP (indicated by the vertical lines). "#" denotes a start-of-sentence marker.

The CBL model divides utterances into chunks, splitting the utterances whenever the

BTP between two words drops below the running average BTP. In the example in Figure 1, the CBL might decide to split the sentence ("did you look at the doggy") into three chunks "did you", "look at", and "the doggy", and store all three in its memory. As it sees more sentences, it would continue to add new chunks and track how often they co-occurred.

The CBL was developed to model children's early speech production and comprehension without appealing to abstract grammatical categories. Specifically, it was designed as an implementation of the hypothesis that children detect and store multi-word chunks using BTP, and also use the stored chunks to parse speech and produce new utterances (see also Arnon and Snider (2010) and Bannard and Matthews (2008)). The model's ability to simulate learning can be measured by first training it on what children hear and then having the model reproduce what children say from the chunks that it learned.

We chose to build on the CBL model because it has successfully accounted for production data in multiple corpora, including developmental datasets. For example: (a) it parsed text better than a shallow parser in three different languages (English, German and French) when using individual words rather than word classes, (b) it was able to recreate up to 60% of child utterance productions in 13 different languages, and (c) it closely replicated data from an artificial grammar learning study (McCauley & Christiansen, 2011; Saffran, 2002). The model has also been able to replicate experimental data on children's multi-word utterance repetitions (Bannard & Matthews, 2008), over-regularization of irregular plural nouns (Arnon & Clark, 2011), and L2-learner speech (see also McCauley & Christiansen, 2014b, 2017). In sum, the CBL model appears to robustly predict the word-chunk patterns in children's speech when given information about what they hear in their input.

Testing for change with age

Following McCauley and colleagues (2011, 2014a, 2019) we tested the CBL model's ability to learn language by checking how well it can reconstruct children's utterances from the chunks discovered in their caregivers' speech. As we are interested in developmental change over the first three years of speech production, we analyzed the model's reconstruction ability with two measures:

- "Uncorrected": The binary (success/fail) reconstruction score originally used by McCauley and colleagues (2011, 2014a, 2019).
- "Corrected": A length-and-repetition-controlled reconstruction score that accounts for the fact that longer utterances have more opportunities for error reconstruction, and for the fact that some child utterances contain repetitions of word(s), making multiple reconstructions match the original utterance.

If BTP is an age-invariant mechanism, it should apply equally well to shorter utterances and longer utterances; the latter of which are more often produced as children get older. We therefore tested for age invariance both with the original binary ("uncorrected") reconstruction score and a new ("corrected") score we proposed to account for utterance length and word repetitions. If we find age-invariance, even while controlling for utterance length and word repetitions, it would strongly suggest that the mechanism is stable over the first three years of speech production and not simply influenced by other factors, e.g., utterance length. If not, it would suggest that use of the mechanism, in fact, changes with age (Bannard et al., 2009; Tomasello, 2005; Yang, 2016).

Predictions

Taking the previous findings as a starting point, we investigated whether the CBL could account for child speech production with equal precision over the first four years of life.

We predicted that:

- The CBL would less accurately generate children's speech productions as they grew older; given the assumption that children gradually learn to abstract over the specific "chunks" they encounter (Bannard et al., 2009; Tomasello, 2005; Yang, 2016) and, therefore, their speech should less often directly mirror their linguistic input at later ages. This finding would indicate that the immediate influence of children's language input statistics on their speech production decreases across development.
- Children will be more likely to use words that are not documented in the caregiver speech as they get older. These words could originate from other sources, such as peer speech (Hoff, 2010; Hoff-Ginsberg & Krueger, 1991; Mannle, Barton, & Tomasello, 1992), or non-recorded caregiver speech (B. C. Roy, Frank, & Roy, 2009).
- Younger children's utterances would be reconstructed well on the basis of recently heard speech alone, whereas older children's utterances would be best constructed when considering a longer period of their historical input. Our reasoning was that older children's increased memory capacity (Bauer, 2005; Gathercole et al., 2004; Wojcik, 2013) allows them to draw on older input more easily in producing speech. If so, the findings would suggest that memory plays a critical role in the use of the same learning mechanism with age.

In sum, we expected to find that the CBL's ability to reconstruct children's speech decreases in-line with a concomitant increase in children's linguistic sophistication; an effect driven by children's use of more abstracted representations, words from other speech sources, and their increased ability to use historical input.

Methods

Model

The CBL model (McCauley & Christiansen, 2011) is an incremental, online computational model of language acquisition, that explores the possibility that infants and children parse their input into (multi-word) chunks during the process of acquiring language.

The model takes transcribed speech as input and divides the transcribed utterances into multi-word chunks. Each utterance begins with a start cue (denoted "#"). The exact placement of a chunk boundary within an utterance is determined by two factors: (1) the backward transitional probability (BTP) between consecutive words in the utterance, and (2) the inventory of already-discovered chunks. All newly discovered chunks are saved into the inventory, alongside the BTPs associated with each chunk. The only information that the model tracks and stores are the discovered chunks, the BTPs between words, and the BTPs between discovered chunks. For example, the model might parse the utterances "I see the doggy" and "did you look at the doggy?" into five different chunks, namely "I", "see", "the doggy", "did you", and "look at" based on the BTPs between these words compared to the average BTP found in the corpus so far.

Child utterance reconstruction task

Once the model has been trained on adult utterances, and thereby has discovered chunks in the adults' speech, we can test whether it closely matches the linguistic structures produced by the children in the same caregiver-child corpus. Following McCauley and Christiansen (2011), we use a child utterance reconstruction task to test whether the chunk statistics present in the adults' utterances are also present in the child's utterances. The model reconstructs the child utterances from the chunks and their related BTPs from the adult's utterances at the same age point. This reconstruction **process, which is slightly different from McCauley and Christiansen's (2011) process,** is done in two steps (see Figure 2). First, a child utterance is converted into an unordered bag-of-chunks containing chunks discovered in the adults' speech, in line with the bag-of-words approach proposed in Chang, Lieven, and Tomasello (2008). Whenever the model encounters a word in the child utterance that is not present in the adult-based chunk inventory, it stops processing that utterance.[1] For instance, in the toy example in Figure 2, the child utterance "look at the doggy" would be decomposed and shuffled into the set of chunks "look at" and "the doggy", which were both discovered in the adults' speech (illustrated in the speech bubble). However, if the utterance were "look at the doggy there", and the model had no chunk for the word "there", then it would reject the utterance. After breaking the utterance into known chunks, the model tries to reconstruct it. During reconstruction, the model begins with the utterance start cue and then follows that initial cue with all the chunks from the bag-of-chunks, ordered by the highest transitional probability in relation to the previously selected one. So, the set of chunks "look at" and "the doggy" would be ordered depending on which chunk had the highest BTP with respect to the utterance start cue ("look at"), followed by the chunk with the highest BTP with respect to "look at" ("the doggy").

Materials and Procedure

As input to the model we used transcripts of 1–2-hour recordings of at-home interaction between six North American children and their caregivers who were recorded approximately every two weeks between ages 1;0 and 4;0 (the Providence corpus; Demuth, Culbertson, and Alter (2006)). We pre-processed the transcripts, which were formatted using CHAT conventions (MacWhinney, 2000), such that the input to the model only contained

[1] McCauley and Christiansen (2011) handle these cases differently**. Our CBL implementation is identical to theirs up to this point. Therefore we also** provide sentence reconstruction scores using their original method in the Supplementary Materials.

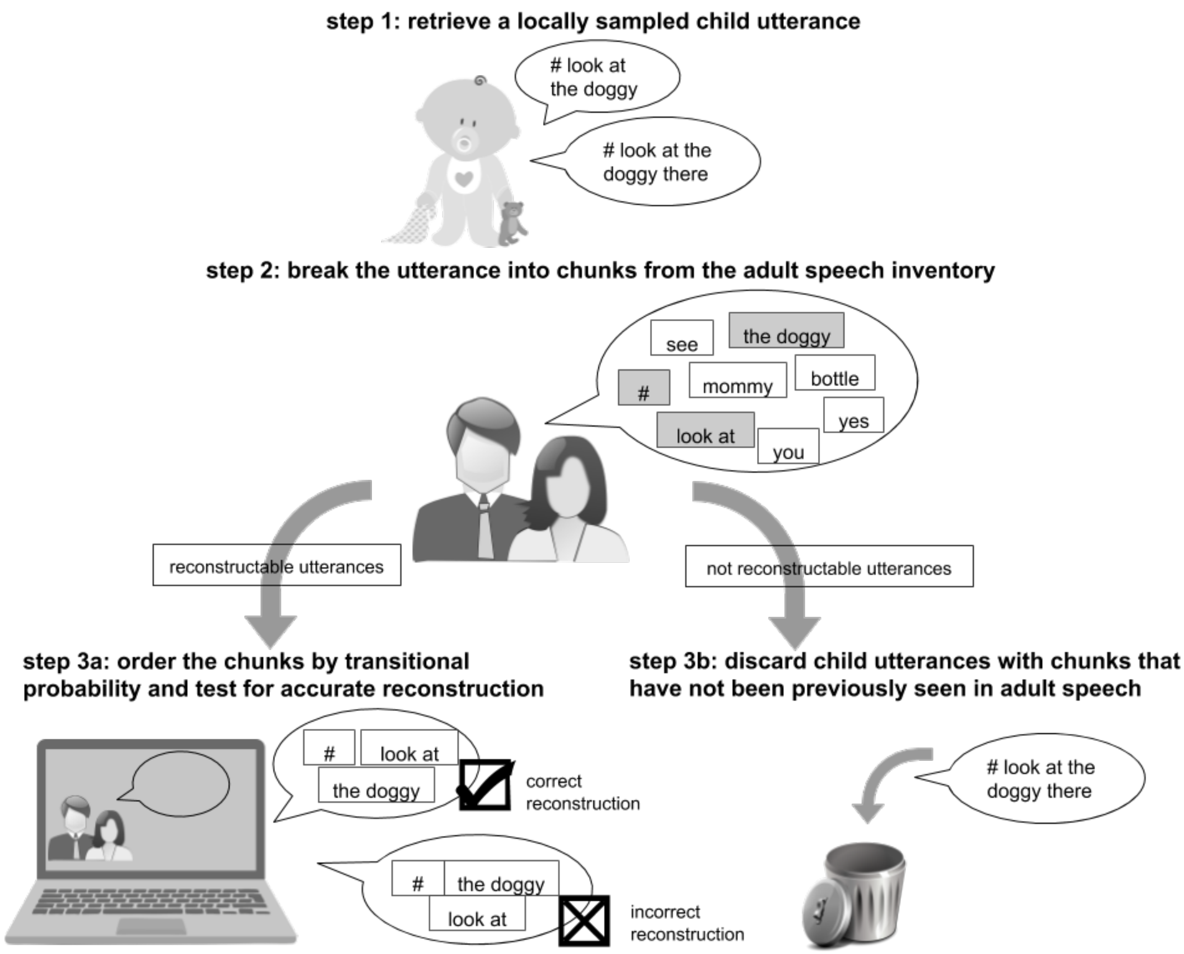

Figure 2. Example of a reconstruction attempt of two child utterances by a toy model. The model is able to reconstruct the first utterance, but it cannot do so with the second utterance, which contains a word ("there") that it has not previously seen.

plain text orthographic transcriptions of what was said.[2] We split the transcripts into two separate files, one with all the caregivers' utterances and one with all the child's utterances. Our pre-processing also added a "#" prefix to the start of each utterance.

The transcripts were sampled at approximate 6-month intervals between ages 1;0 and

[2]All punctuation marks, grammatical notes, omitted word annotations, shortenings, and assimilations were removed from the utterances, such that only the text representing the spoken words of the utterance remained.

4;0. We used two different sampling methods: a local data sampling method and a cumulative data sampling method. With the local data sampling method we selected data within a two-month interval around each age point. For example, for age point 1;6 we selected transcripts in which the child was between 1;5.0 and 1;6.31 years of age. This method led to ~800–4000 caregiver utterances at each age point. By design, the local sampling method focuses the model's training solely on *recent* linguistic input so that, when it tries to reconstruct children's utterances, the result is a test of how closely their current speech environment can account for what they say.

In contrast, the cumulative sampling method focuses the model's training on all previously heard linguistic input so that, when it tries to reconstruct children's utterances, the result is a test of how closely their current *and* previous speech environments can account for what they say. For the cumulative sample we selected data for each age point by taking all the available transcripts up to that age point. For example, for age 1;6 we selected all transcripts in which the child was 1;6 or younger. This method led to ~800–60,000 caregiver utterances across the different age points, with the number of caregiver utterances increasing (i.e., accumulating) with child age. As a consequence, the cumulative sample always contained more caregiver utterances than the local sample, except at age 1;0, the first sampled age point.

While we used two different sampling methods for training the model on adult data, all child utterances used for the reconstruction task were retrieved using the local sampling method for that particular age point. In other words, we only reconstructed the child utterances local to each tested age, regardless of the training strategy.

Analysis

We modeled two primary scores related to utterance reconstruction: the uncorrected (binary: success/fail) reconstruction score used by McCauley and colleagues (2011, 2014a, 2019) and the corrected reconstruction score we introduce in the current paper. The uncorrected reconstruction score (1: success, 0: fail) was computed for all child utterances that could be decomposed into previously seen chunks (see step 3a in Figure 2). The corrected reconstruction score (defined below) was computed for the same set of utterances. We additionally included a third analysis: the likelihood that an utterance contains previously unseen words which, by our version of the CBL, cannot be reconstructed (see step 3b in Figure Figure 2).

We used mixed-effects regression to analyze the effect of child age on both of the reconstruction scores and whether utterances contained previously unseen words. All mixed-effects models included child age as a fixed effect and by-child random intercepts with random slopes of child age. The mixed-effects model of utterances with previously unseen words also included the number of words in the utterance, as explained below. By default, child age was modeled in years (1–4) so that the intercept reflects a developmental trajectory beginning at age 0. However, for the model of corrected reconstruction accuracy, we centered child age on zero in order to test whether the CBL's reconstruction score was above chance (i.e., contributing to accuracy significantly above what would be expected by chance) at the average age (2;6 years) in the dataset.

All analyses were conducted using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) and all figures were generated with the `ggplot2` package in `R` (R Core Team, 2014; Wickham, 2009). All code used to create the model and analyze its output is available at https://osf.io/ca8ts/?view_only=daaa1bcc71654842b0d70efe785a26b9. Before turning to the main results we briefly describe the corrected reconstruction score and the analysis of previously unseen words in more detail.

Corrected reconstruction accuracy

The corrected, length-and-repetition-controlled reconstruction score is a function of three factors: (a) whether the model successfully reconstructed the child utterance or not, (b) the number of chunks used to reconstruct the utterance, and (c) the number of duplicate chunks involved in the reconstruction. By taking the number of chunks into account, this reconstruction score compensates for the fact that successful reconstruction is less likely for longer utterances. When an utterance contains duplicate chunks, the exact ordering of those duplicate chunks does not influence the correctness of the reconstruction. For example, if the utterance "I wanna, I wanna" is decomposed into the two chunks "I wanna" and "I wanna", it does not matter which of the two "I wanna" chunks is placed first when calculating the reconstruction accuracy of the utterance. Thus, utterances containing duplicate chunks are more likely to be reconstructed by chance alone than utterances with the same number of chunks with no duplicates. **Note that here we detecting duplicate *chunks* in the utterance rather than duplicate *words*. At this post-training stage, the model is only able to parse the utterance into chunks; that is the relevant unit over which duplication may affect reconstruction accuracy.**

An utterance that is decomposed into N unique chunks can be reconstructed in $N!$ different orders. Hence, the probability of obtaining the correct order of N unique chunks merely by chance equals $1/N!$. When we take into account that chunks can be repeated within an utterance, chance level equals $(n_1!n_2! \dots n_k!)/N!$ where N is the total number of chunks in the utterance, and $n_1, \dots, n_k$ are the number of times a chunk is repeated for each of the k unique chunks found in the utterance.

When probability of reconstruction was lower, we scored a correctly reconstructed utterance higher. We assigned a score of $-\log(chance)$ for each correct reconstruction and $\log(1 - chance)$ for each incorrectly reconstructed utterance. In layman's terms, this means that successfully reconstructed utterances were scored positively, but were weighed relative to the number of chunks and the number of repetitions they had, such that reconstructions of long utterances were given higher scores than reconstructions of short utterances. Along the same lines, incorrectly reconstructed utterances were scored negatively and were also weighed relative to the number of chunks they had, such that incorrect reconstructions of long utterances were given higher (i.e., less negative) scores than incorrect reconstructions of short utterances.

To illustrate the corrected scoring method, let's compare two three-chunk utterances, one of which contains a duplicate chunk: “I wanna I wanna see” (chunks: “I wanna”, “I wanna”, “see”) and “I wanna see that” (chunks: “I wanna”, “see”, “that”). For the first utterance, chance level equals $(2! \times 1!)/(3!)$: The numerator is determined by the number of times each unique chunk is used, so because “I wanna” occurs two times and “see” occurs once, that is $2! \times 1!$. The denominator is determined by the factorial of total number of chunks (here: $3! = 3 \times 2 \times 1$). The resulting chance level is then $2/6$. For the second utterance, chance level equals $(1! \times 1! \times 1!)/(3!)$: The numerator is equal to $1! \times 1! \times 1!$ here because all chunks occur only once in the utterance. The denominator is the same as for the first utterance as the total number of chunks in the utterance is the same. Here, the resulting chance level is $1/6$. If the utterances are reconstructed correctly, the score is computed by $-\log(chance)$. So, the first utterance would get a positive score of $-\log(chance) = -\log(2/6) \approx 1.098$ and the second utterance would get a higher positive score of $-\log(chance) = -\log(1/6) \approx 1.791$ for increased reconstruction difficulty. If the utterances are reconstructed incorrectly, the score is computed by $\log(1 - chance)$. Thus, the first utterance would get a negative score of $\log(1 - chance) = \log(1 - (2/6)) \approx -0.405$ and the second utterance would get a less negative score of $\log(1 - chance) = \log(1 - (1/6)) \approx -0.182$.

Previously unseen words

Our third analysis focused on whether or not each child utterance contained words that were not previously seen, that is, not present in the trained adult-speech chunk inventory. Each utterance was coded for unseen words with a binary value (1: at least one unseen word, 0: no unseen words). Because there is no straightforward way to establish a baseline probability for a word not being present in the adult chunk inventory, we could not incorporate a precise baseline for the probability of encountering unseen chunks into the scoring. Instead, we used the number of words in each utterance as an additional factor in the analysis on the assumption that longer utterances are, in general, more likely to contain unseen words.

Results

Uncorrected reconstruction accuracy

The uncorrected score of accurate utterance reconstruction (McCauley & Christiansen, 2011, 2014a) showed that model's average percentage of correctly reconstructed utterances across children and age points was similar for the locally and cumulatively sampled speech (local: mean = 65.4%, range across children = 59.9%–70.3%; cumulative: mean = 59.9%, range across children = 53.1%–68.2%). This is similar to, or slightly higher than, results reported by McCauley and Christiansen (2011) who found an average percentage of correctly reconstructed utterances of 59.8% over 13 typologically different languages with a mean age range of 1;8–3;6 years. Additionally, McCauley and Christiansen (2019) reported an average reconstruction percentage of 55.3% for 160 single-child corpora of 29 typologically different languages, including a performance of 58.5% for 43 English single-child corpora with a mean age range of 1;11–3;10.

In our statistical models of the uncorrected reconstruction accuracy[3], we first analyzed the CBL model's performance when it was trained on locally sampled caregiver speech. The number of correctly reconstructed utterances decreased with age ($b = -0.805, SE = 0.180, p < 0.001$); over time the BTP statistics present in the caregivers' speech were less reflected in the child's own speech (Figure 3, left panel).

We then tested the model's performance when it was trained with a cumulative sample of caregiver speech, rather than just a local sample. As before, the number of correctly reconstructed utterances decreased with child age ($b = -0.821, SE = 0.146, p < 0.001$; Figure 3, right panel). These results indicate age-variance for the SL mechanism; its utility for modeling children's utterances changes with age.

Importantly, however, the length of the child utterances varied quite a lot (range = 1–44 words long; mean = 2.8, median = 2), and some of them contained repetitions of chunks (e.g., "I wanna, I wanna"), both of which influence the baseline likelihood of accurate reconstruction. Utterances from older children tended to contain more words than utterances from younger children (Figure 4, left panel). As a consequence, on average, utterances from older children are systematically less likely to be correctly reconstructed by chance, contributing to the decrease in the CBL's overall performance with age. Additionally, the percentage of child utterances that contained duplicate chunks decreased over time (Figure 4, right panel). Utterances with duplicate chunks had a higher baseline probability of being accurately reconstructed by the model. So again, on average, utterances from older children were systematically more difficult, contributing to the age-related decrease in uncorrected reconstruction scores.

[3]**accuracy ~ age + (age|child), family = binomial(link = "logit").**

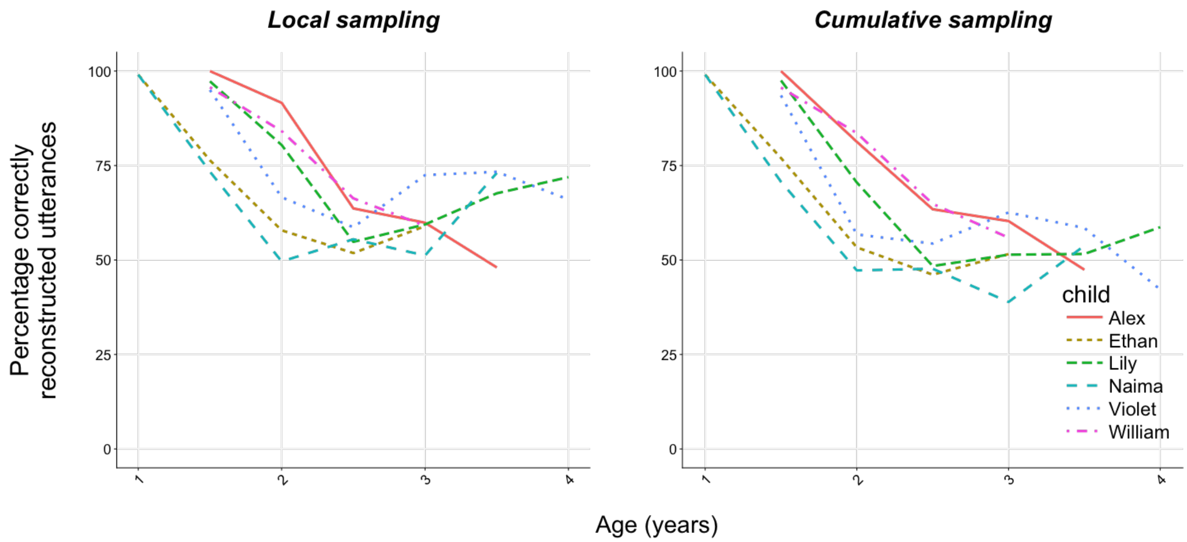

Figure 3. Percentage of correctly reconstructed utterances across the age range, using local (left) and cumulative (right) sampling.

Corrected reconstruction accuracy

Next, we used our corrected reconstruction score to assess the model's reconstruction accuracy while controlling for utterance length and the use of duplicate chunks. The score weighs whether each utterance was accurately reconstructed against its chance level of reconstruction, depending on the total number of chunks and number of duplicate chunks it contains. The model's average reconstruction score across children and age points was similar for the locally and cumulatively sampled speech (local: mean = 0.10, SE = 0.01; cumulative: mean = 0.06, SE = 0.01). Note again that one aim of this analysis was to test whether the corrected reconstruction score was above chance—here represented by a score of zero—so in the statistical models we centered child age on zero so that the estimation would reflect the difference from zero for the average age in our sample (2;6 years).[4]

Again, we first analyzed the model's performance when it was trained on locally sampled caregiver speech. We found a significant positive intercept

[4]**accuracy ~ age + (age|child).**

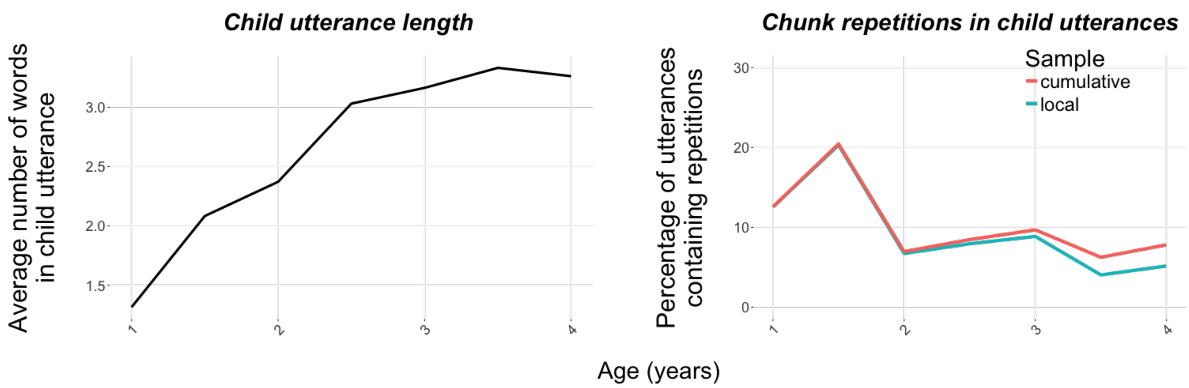

Figure 4. Children's utterances increased in length (number of words) with age (left) while simultaneously decreasing in the number of duplicate chunks used (right).

$(b = 0.11, SE = 0.02, t = 5.064)$ and no significant change across age $(b = 0.030, SE = 0.018, t = 1.681)$; the BTP statistics from the caregivers' speech were consistently reflected in the child's own speech (Figure 5, left panel).

As before, we created a parallel set of analyses to test the model's performance when it was trained with a cumulative sample of caregiver speech. We again found a significant positive intercept $(b = 0.06, SE = 0.010, t = 6.238)$ and that accuracy did not change significantly across age ($b = 0.02, SE = 0.013, t = 1.590$; Figure 5, right panel).

In sum, contrary to the uncorrected reconstruction accuracy analysis, these corrected reconstruction score results indicate age-invariance for the SL mechanism. In addition, the model performed significantly above chance level in both the local and cumulative sampling contexts.

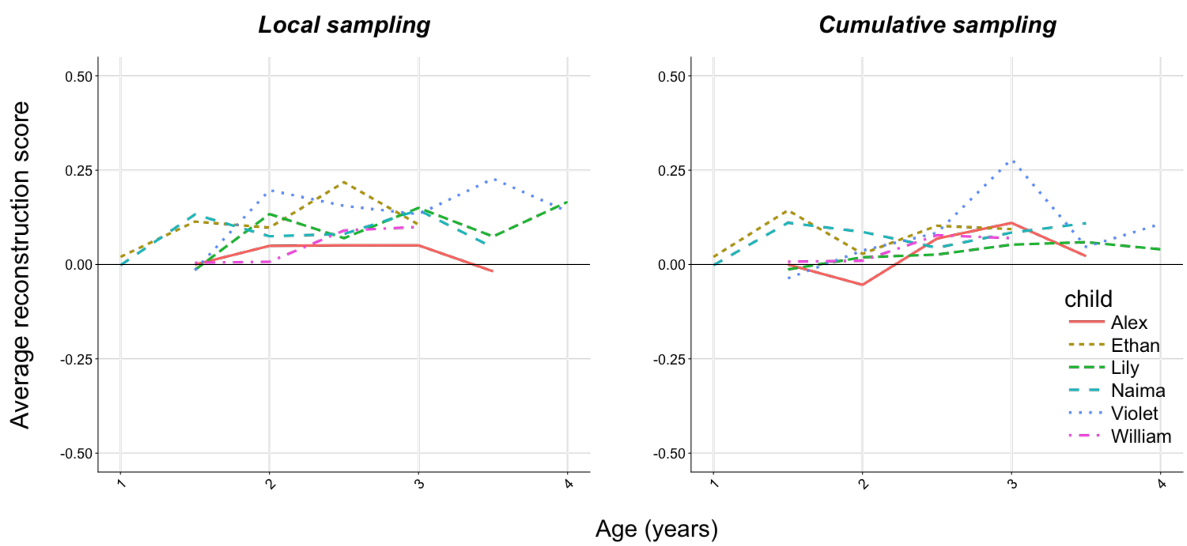

Figure 5. Corrected reconstruction scores across the age range, using local (left) and cumulative (right) sampling.

Children's use of unseen words

Finally, we modeled whether an utterance had a previously unseen word or not (1: at least one unseen word, 0: no unseen words).[5] Using local sampling the number of child utterances that contained one or more previously unseen words increased with age ($b = 0.364, SE = 0.093, p < 0.001$; Figure 6, left panel). Unsurprisingly, longer utterances were also significantly more likely to have previously unseen words in them ($b = 0.170, SE = 0.005, p < 0.001$). By taking a longer history of linguistic input into account (i.e., by using cumulative sampling), we expected to see a smaller increase in previously unseen words with age. We indeed found that utterances containing previously unseen words became *less likely* with age ($b = -0.139, SE = 0.068, p < 0.05$; Figure 6, right panel). Also, as before, previously unseen words were more likely to occur in longer utterances ($b = 0.105, SE = 0.006, p < 0.001$).

[5] **has_unseen_words ~ age + (age|child) + num_words_in_utt, family = binomial(link = "logit")**

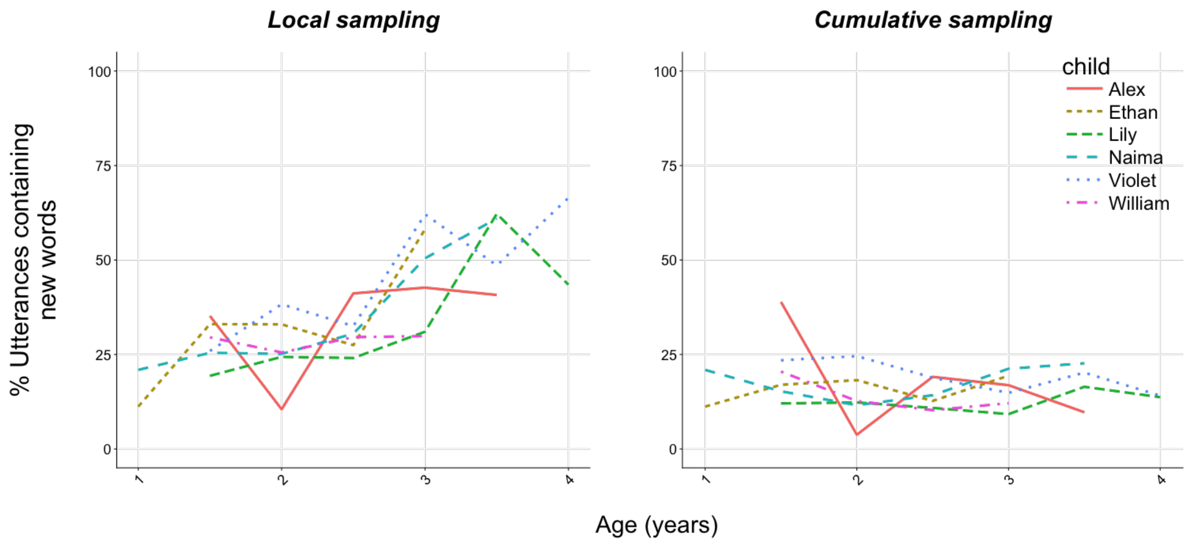

Figure 6. Number of utterances with previously unseen words across the age range, using local (left) and cumulative (right) sampling.

Discussion

Our primary research question (as raised by, e.g., Arciuli & Simpson, 2011; Raviv & Arnon, 2018; J. R. Saffran et al., 1997; Shufaniya & Arnon, 2018) was whether the CBL is able to predict children's speech productions with stable accuracy throughout development. We tested the model using both the original measure of accuracy as well as a new measure that takes into account utterance length and duplicate chunks in the utterance, which can make accurate reconstruction less likely (length) or more likely (duplicates). Using this corrected measure, we found that there was no significant change in the use of BTP with age. Notably, the CBL was able to construct utterances at above-chance levels despite these changes with age. Overall, our findings support the view that BTP is an age-invariant learning mechanism for speech production. In fact, the positive, but non-significant coefficients for the effect of age on corrected reconstruction accuracy indicate that, the CBL is, at least, not getting worse at reconstructing children's utterances with age.

Different words at different ages

We also analyzed the number of utterances with previously unseen words in them, arguing that older children's increased memory capacity (Bauer, 2005; Gathercole et al., 2004; Wojcik, 2013) would possibly allow them to draw upon older input more easily in producing speech. Indeed, we found an increase in the number of utterances containing previously unseen words with age in the local sample but a decrease when taking their longer linguistic history into account. The change in word usage we find here could be partly due to a change in linguistic input not captured in the transcripts. The corpus we used is relatively dense: multi-hour at-home recordings made approximately every two weeks for 2–3 years. However, this corpus still only contained a small fraction of what each child heard during the represented periods of time (i.e., 2 hours of ~200 waking hours in a fortnight). Non-recorded caregiver speech may contribute an increasing amount of lexical diversity. Consider, for example, that input from peers containing different lexical items could have increased as children became old enough to independently socialize with other children or attend daycare or preschool (Hoff, 2010; Hoff-Ginsberg & Krueger, 1991; Mannle et al., 1992), which may help to account for the increased presence of words not found in the caregiver's speech. This problem is difficult to address directly since, even with cutting-edge tools and significant supporting resources, it is still nearly impossible to collect and transcribe a child's complete language environment (Casillas & Cristia, under review; B. C. Roy et al., 2009). This effect could instead be simulated in future work by feeding speech from other children or adults into the model to mimic speech from peers and other caregivers. That said, our results showed that the likelihood of previously unseen words actually decreased with age for the cumulative sample, suggesting that the "missing" words *are* present in caregiver speech, just not always in the recently recorded input.

Additionally, an improvement in memory capacity with age provides a potential explanation for the current findings. Throughout childhood, including the first few years, SL-relevant cortical regions continue maturing (Casey et al., 2000; Diamond, 2002; Rodríguez-Fornells et al., 2009; Uylings, 2006) with concurrent increases in long-term memory (Bauer, 2005; Wojcik, 2013), working memory, and speed of processing (Gathercole et al., 2004; Kail, 1991). By ages three and four, the children in the current study may have been able to much more reliably draw upon information they were exposed to in the more distant past. If so, we would expect no significant increase in the use of previously unheard words as children get older with the cumulative sampling method—consistent with what we found here (Figure 6, right panel). This pattern of results indicates that children's developing memory could play an important role in the way they use environmental input statistics over age.

Abstraction and complex utterances

Our findings are not consistent with a representational shift toward abstraction during the early language learning process. For instance, if children schematized their constructions or switching to rule-based representations (Bannard et al., 2009; Tomasello, 2005; Yang, 2016), we would expect a decrease in reconstruction accuracy over time, given that the CBL's reconstructions are limited to the immediate statistics of the child's language environment. In contrast, we saw that the model's ability to reconstruct child utterances from caregivers' speech was age-invariant when taking into account utterance length and chunk duplicates. These results do fall in line with SL theories proposing that the mechanisms for processing, storing, and deploying information stay the constant over age, even though SL behavior on the surface might seen to change over time (e.g., Misyak et al., 2012).

As the CBL model only employs a single, simple mechanism for creating and tracking linguistic units, it is impressive that it performs at above-chance levels when accounting for children's speech productions in the first few years. If the mechanism is truly age-invariant, it should be able to handle both young children's speech and adults' speech; here we see that it handles the developing linguistic inventory of children ages 1;0 to 4;0, during which time children's utterances come much more sophisticated and much closer to adult-like form.

Going beyond the scope of this paper, a next step would be to explore how the CBL could be modified to augment its performance, particularly on more complex utterances. For example, the CBL model does not include the use of semantics when dividing the caregivers' speech into chunks or when reconstructing the child utterances. However, the meaning of what both caregivers and child are trying to convey plays a fundamental role in selecting words from the lexicon and in constructing utterances—they are interacting, and not just producing speech. The same set of words, ordered in different ways, can have entirely different meanings (e.g., "the dog bites the man" vs. "the man bites the dog"). Additionally, the CBL currently works on text-only transcriptions of conversations, but speech prosody could potentially critically change how children detect chunks. Prosodic structures within an utterance highlight syntactic structures and help to distinguish between pragmatic intentions, for example, distinguishing between questions, imperatives, and statements (e.g., Bernard & Gervain, 2012; Speer & Ito, 2009). Ideally, the CBL model would also be tested on a (more) complete corpus of what children hear in the first few years to further investigate the origins of the "previously unseen" words in children's utterances; though we appreciate that densely sampled and transcribed collections of audio recordings are extremely costly to create (Casillas & Cristia, under review; B. C. Roy et al., 2009).

Conclusion

In this study, we investigated whether the CBL model—a computational learner using one SL mechanism (BTP)—could account for children's speech production with equal accuracy across ages 1;0 to 4;0 given information about their speech input. The model's ability to reconstruct children's utterances remained stable with age when controlling for utterance length and duplicate chunks, both when taking into account recent and cumulative linguistic experience. These findings suggest that this particular mechanism for segmenting and tracking chunks of speech may be age-invariant (Raviv & Arnon, 2018; Shufaniya & Arnon, 2018). A rich and growing literature on SL in development has demonstrated that similar mechanisms can account for much of children's early language behaviors; knowing whether the use of these mechanisms changes as children get older is a crucial piece of this puzzle. To explore this topic further, future work will have to address additional cues to linguistic structure and meaning, the density of data needed to get reliable input estimates, and the interaction of SL with other developing skills that also impact language learning.

Acknowledgements

We owe big thanks to Rebecca L. A. Frost and the MPI for Psycholinguistics' First Language Development group for insightful comments on earlier versions of this paper. This work was supported by an IMPRS fellowship awarded to IR and a Veni Innovational Research grant to MC (275-89-033).

References

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science*, *14*(3), 464–473.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth–Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, *92*, 265–280.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284–17289.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bauer, P. J. (2005). Developments in declarative memory: Decreasing susceptibility to storage failure over the second year of life. *Psychological Science*, *16*(1), 41–47.

Bernard, C., & Gervain, J. (2012). Prosodic cues to word order: What level of representation? *Frontiers in Psychology*, *3*, 451.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132.

Casey, B. J., Giedd, J. N., & Thomas, K. M. (2000). Structural and functional brain development and its relation to cognitive development. *Biological Psychology*, *54*(1-3), 241–257.

Casillas, M., & Cristia, A. (under review). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 154–159.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, *9*(3), 198–213.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *31*(1), 24–39.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, *49*(2), 137–173. doi:doi:10.1177/00238309060490020201

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. Stuss & R. Knights (Eds.), *Principles of frontal lobe function* (pp. 466–503). New York: Oxford University Press.

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *The Quarterly Journal of Experimental Psychology*, *64*(5), 1021–1040.

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: High frequency words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, Advance online publication.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177–190.

Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. *First Language*, *30*(3-4), 461–472.

Hoff-Ginsberg, E., & Krueger, W. M. (1991). Older siblings as conversational partners. *Merrill-Palmer Quarterly*, *37*(3), 465–481.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8-and 11-month-olds. *Infancy*, *14*(1), 2–18.

Jost, E., & Christiansen, M. H. (2016). Statistical learning as a domain-general mechanism of entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 227–244). Washington D.C.: Mouton de Gruyter.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, *109*(3), 490–501.

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in snfant speech segmentation: Achieving the lexical shift. *Infancy*, *23*(6), 770–794.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247–1252.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.).

Psychology Press.

Mannle, S., Barton, M., & Tomasello, M. (1992). Two-year-olds' conversations with their mothers and preschool-aged siblings. *First Language*, *12*(34), 57–71.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1619–1624.

McCauley, S. M., & Christiansen, M. H. (2014a). Acquiring formulaic language: A computational model. *The Mental Lexicon*, *9*(3), 419–436.

McCauley, S. M., & Christiansen, M. H. (2014b). Reappraising lexical specificity in children's early syntactic combinations. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 1000–1005.

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*(3), 637–652.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Pyschological Review*, *126*(1), 1–51.

Misyak, J. B., Goldstein, M. H., & Christiansen, M. H. (2012). Statistical-sequential learning in development. *Statistical Learning and Language Acquisition*, 13–54.

Monroy, C. D., Gerson, S. A., & Hunnius, S. (2017). Toddlers' action prediction: Statistical learning of continuous action sequences. *Journal of Experimental Child Psychology*, *157*, 14–28.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning.

Cognition, *126*(2), 268–284.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–247.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*(4), e12593.

Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & Diego-Balaguer, R. de. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3711–3735.

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 2106–2111.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172–196.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181–203.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science*, *42*(8), 3100–3115.

Slone, L. K., & Johnson, S. P. (2015). Infants' statistical learning: 2-and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, *133*, 47–56.

Speer, S. R., & Ito, K. (2009). Prosody in first language acquisition–Acquiring intonation as a tool to organize information in conversation. *Language and Linguistics Compass*, *3*(1), 90–110.

StClair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*(3), 341–360.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, 21.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition* (1st ed.). Harvard University Press.

Tomasello, M. (2008). Acquiring linguistic constructions. In Damon W, R. Lerner, D. Kuhn, & R. Siegler (Eds.), *Child and adolescent development* (pp. 263–297). New York:

Wiley.

Uylings, H. B. M. (2006). Development of the human cortex and the concept of "critical" or "sensitive" periods. *Language Learning*, *56*(1), 59–90.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer Publishing Company, Incorporated.

Wojcik, E. H. (2013). Remembering new words: Integrating early memory development into word learning. *Frontiers in Psychology*, *4*, 151.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language* (1st ed.). MIT Press.