Journal of Child Language (2025), 1–40
doi:10.1017/S0305000925100159

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Simulating procedural discovery in early language acquisition: Domain-general cognition with contextual learning

Yang Ji, Jacolien van Rij and Niels Taatgen

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, The Netherlands
Corresponding author: Yang Ji; Email: y.ji@rug.nl

(Received 05 April 2024; revised 27 June 2025; accepted 03 July 2025)

Abstract

We present a simulation study based on a cognitive architecture that unifies various early language acquisition phenomena in laboratory and naturalistic settings. The model adaptively learns procedures through trial-and-error using general-purpose operators, guided by learned contextual associations to optimise future performance. For laboratory-based studies, simulated preferential focusing explains the delayed behavioural onset of statistical learning and the possible age-related decrease in algebraic processing. These findings suggest a link to continuous, implicit learning rather than explicit strategy acquisition. Moreover, procedures are not static but can evolve over time, and multiple plausible procedures may emerge for a given task. Besides, the same model provides a proof-of-concept for word-level phonological learning from naturalistic infant-directed speech, demonstrating how age-related processing efficiency may influence learning trajectories implicated in typical and atypical early language development. Furthermore, the artile discusses the broader implications for modelling other aspects of real-world language acquisition.

Keywords: computational model; cognitive architecture; learning mechanism

1. Introduction

Young children are remarkable language learners. They effortlessly acquire syntactic rules and lexical forms of their native language, even with limited exposure. This ability has long intrigued linguists and developmental psychologists (for reviews, see Kuhl, 2004; Saffran & Kirkham, 2018). Two prominent theories approach this ability from different angles. The statistical learning perspective suggests that young children can learn and differentiate between lexical forms based on syllable features (e.g., Estes et al., 2007; Saffran et al., 1996; Saffran & Kirkham, 2018). Alternatively, the algebraic theory posits that children learn to generalise over variable lexical forms to acquire abstract syntactic patterns (e.g., Frank & Tenenbaum, 2011; Marcus et al., 1999). Both statistical learning and algebraic theories make valuable contributions to our understanding of language

acquisition but focus exclusively on specific aspects of the overall language acquisition problem. This article presents a simulation study that aims to explain simple forms of lexical and syntactic learning within a unified computational theory. Our study aligns with the aim to "model more phenomena, integrate across linguistic levels" and the theoretical prospect that clarifies the "domain-general underpinnings of more aspects of languages" or "the impact of cognitive processing" (e.g., see Benders & Blom, 2023).

We implemented our computational simulations of early linguistic acquisition phenomena within a cognitive architecture. Cognitive architectures offer a computational framework for simulating a wide range of tasks within a unified framework (Anderson, 2007; Newell, 1973). A cognitive architecture is a general theory of cognition that encompasses, but is not limited to, linguistic processing. Cognitive architectures, such as SOAR and ACT-R (Anderson, 2007; Laird et al., 1987), share a common structure (for a review, see Laird et al., 2017), including memory hubs (modules) that store information and procedural knowledge that specifies the possible movements and comparisons of information content between modules. Generally, procedural knowledge is represented as a sequence of context-dependent or conditional *production rules* (e.g., given an if-condition is true, a specific then-action processing follows) that process information and implement strategies (expounded later; see Figure 2). The simulations of the cognitive processing in these tasks yield behavioural results that can be compared with empirical findings. This approach has already facilitated a comprehensive understanding of human cognition across various cognitive domains (see Kotseruba & Tsotsos, 2018, for a review). In this study, we then apply the general cognitive framework of cognitive architecture to explore its applicability to various levels of interest in early linguistic processing.

Despite their domain-general structure, cognitive architectures often lack cross-domain learning mechanisms. This is because, in procedural knowledge, the contexts (if-conditions) of each production rule are manually defined. This limitation hinders the ability of general-purpose processing (then-actions) to adapt to other contexts in a dynamic environment. As Taatgen (2017) emphasises, the reliance on rigid, pre-programmed production rule sequences restricts cognitive architectures from adjusting to changing conditions and introduces inconsistencies among different models. This reliance on pre-programmed production rule sequences poses particular challenges in modeling the learning processes of young children, who often acquire new information processing steps through trial and error, without explicit instructions. For instance, young children may initially develop multiple equally plausible interpretations of a single multisyllabic pattern on a trial-by-trial basis (e.g., Gerken, 2006, 2010), similar to how adults may exhibit multiple strategies when faced with uninstructed tasks, such as solving a Tower of Hanoi problem (e.g., Simon & Newell, 1971). Therefore, the explanation of early linguistic acquisition also requires incorporating a learning mechanism that allows an undifferentiated cognitive architecture to adapt to a variety of task environments.

This article presents a unified computational framework for early linguistic theories, based on the primitive information processing elements architecture (PRIMs, Taatgen, 2013) driven by an underlying contextual learning mechanism. Unlike traditional cognitive architectures, PRIMs decomposes production rule sequences into a collection of general-purpose processing elements (then-actions with minimal conditions) and allows the model to learn context-binding (if-conditions) associatively through interaction with the task environment. Consequently, the model does not have any preconceived rule-like anticipations about the procedural steps to follow. Instead, procedural knowledge is gradually discovered through trial and error. This allows the PRIMs architecture to be more open in terms of the procedural knowledge that may emerge from processing a given task. The first research question of this study specifically examines the types of

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

procedural knowledge that may be constructed from processing various tasks traditionally hypothesised to be related to either lexical or syntactic abilities.

Additionally, procedural learning in PRIMs is now guided by a contextual learning mechanism, representing task experience as a set of associations between relevant model contexts and applied elements. This contrasts with the explicitly all-or-none rule-based procedural knowledge about a task. The second research question examines whether the empirical observation of preferential focusing on learned or novel tasks indicates all-or-none strategy acquisition or simply the differentiation of patterns due to continuous implicit procedural learning.

In the next sections, we will first outline the linguistic tasks that are the primary focus of the simulations carried out in this study. Subsequently, we will delve into the details of the PRIMs cognitive architecture, which is utilised to simulate the linguistic phenomena of interest.

2. Making sense of early linguistic phenomena

Previously, lab-based studies have focused on statistical learning and algebraic tasks in isolation, without considering them as phenomena along a lexical-syntactic continuum (see Bates, 1979). In the following, we will briefly outline these studies and their subsequent extensions. Next, we will introduce an alternative elemental strategy perspective that seeks to unify these diverse perspectives. Finally, we will critically examine the question of strategy learning in light of the limitations of relying solely on indirect measures of preferential focusing dynamics.

2.1 Lexical statistical learning

The first series of tasks explores how young children acquire new lexical forms, as seen in the study of Saffran et al. (1996) with 8-month olds. To control for prior experience, the study employs pseudowords composed of fixed syllables (e.g., *X-Y-Z*, where each uppercase letter denotes a concrete syllable). The words are presented as a continuous syllable stream without breaks. Following familiarisation with specific word patterns, the children's ability to discriminate between these patterns and novel alternatives, namely nonwords/part-words, is assessed. The results reveal that young children show increased attention to novel, untrained patterns. This preference for novelty is interpreted as evidence for sufficient learning of the word pattern, allowing them to differentiate it from alternatives.

The statistical learning perspective interprets the empirical phenomenon by attributing it to the influence of the external task environment, rather than focusing on the learning capabilities of the cognitive system (Saffran & Kirkham, 2018). The initial theory is specifically concerned with the transitional probability of adjacent syllables, which refers to the likelihood of encountering a particular next syllable given the preceding syllable (Saffran et al., 1996). A high transitional probability indicates that adjacent syllables are likely to occur together within a word form, whereas a low transitional probability implies that the syllables either do not co-occur or only partially overlap. Lower transitional probability can occur in completely novel words or at word boundaries where a syllable may be followed by a variety of other words.

Another subsequent study conducted within the same paradigm poses the question of whether children solely learn the immediate transitional probabilities between adjacent

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

syllables or if they also acquire multisyllabic phonological forms that enable them to learn meaningful lexical referents (17-month-olds, Estes et al., 2007). In this study, the children underwent a similar familiarisation phase as in Saffran et al. (1996), but an additional object-label learning phase was introduced. In this phase, both trained words and non-words/part-words (labels) were paired with images (objects). They found that young children only exhibited dishabituation when the paired image was changed in the image-word condition. This indicates that they formed a link between the trained word labels and the images during the subsequent object-label learning phase, but not between untrained non-word/part-word labels and the images. The results suggest that young children's statistical learning extends beyond simple syllable order, encompassing the ability to acquire and utilise multisyllabic forms for referencing purposes (Estes et al., 2007).

Until now, statistical learning has been extensively studied for nearly three decades and is considered a fundamental aspect in language acquisition (e.g., word learning Bergmann & Cristia, 2015; Saffran & Kirkham, 2018). Meta-analyses demonstrate its robustness across various testing conditions (medium effect, in infants), but the factors influencing its learning effects remain unclear (Isbilen & Christiansen, 2022). For instance, several factors (e.g., age range, stimulus format and pattern number, training and test length, and testing method) have shown no effect on statistical learning. Most surprisingly, the strength of transitional probability is also among these null moderators. In contrast, the presence of additional cues – including social, prosodic, and additional visual or auditory cues – appears to be the only reliable moderator, enabling young children to differentiate informative input features from others. Taken together, the specific mechanisms underlying statistical learning remain to be clarified.

2.2 Syntactic algebraic processing

A second series of tasks explores how 7-month-old children process simple syntactic structures, as seen in Marcus et al. (1999). While also utilising trisyllabic patterns, these studies depart from fixed syllables by employing syllable classes with a repeating structure (e.g., repetition of syllable class *a* in *a-b-a*, where each lowercase letter denotes a variable syllabic type). These patterns mimic syntactic structures. After familiarisation, children's ability to recognise syntactically consistent and inconsistent trisyllabic patterns is tested. Crucially, the syllables previously instantiated in the training phase and their transitional probabilities offer no clues for distinguishing the test patterns, since these syllables are not used in the test phase. Despite this, young children still exhibit similar preferential attention to novel syntactic patterns, demonstrating their ability to differentiate familiar from novel structures.

While rule-based algebraic perspectives emphasise internal cognitive processes for explaining the phenomenon (e.g. Pinker, 1999), they acknowledge that learning is still necessary. This is evident in the training phase, where children need to distinguish between the test conditions that are consistent and inconsistent in type (e.g., between *c-d-c* and *c-d-d* after training *a-b-a*). To address the learning aspect, Frank & Tenenbaum (2011) proposed a rule-based Bayesian model where pre-existing rules are differentiated during the training phase. Further refining this approach, Frank et al. (2016) reframed rules as processing sequences within a cognitive architecture, suggesting they are acquired through interaction with task contexts.

Moreover, the ability to flexibly acquire processing sequences implies that young children may also explore alternative processing sequences (strategies) to process the task

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

presentation, which can lead to a different interpretation of the task. Gerken and colleagues' work shows that the original task of Marcus et al. (1999) allows for at least two strategies (Gerken, 2006, 2010). One strategy is the assumed abstract, rule-based processing sequence. Additionally, Gerken (2006) proposes the emergence of a more lexical strategy, which focuses on particular syllables consistently appearing in a fixed position (e.g., *X* always appears as the middle token in *a-X-a*). Gerken's experiments explored how young children adapt their processing strategies based on the task environment. Children first learned to distinguish between different types of generalised patterns. This training successfully enabled them to differentiate generalised test patterns later. However, when trained on specific patterns, they no longer distinguished generalised test patterns but instead successfully distinguished specific test patterns (Gerken, 2006, 2010). This demonstrates that young children can flexibly adjust their processing strategies. In this case, when the task shifted to a more lexical focus, they shifted and adopted lexical strategies. These findings raise a question: would children stick to a generalised strategy to form a syntactic interpretation, or would they instead process the task by switching to a lexical interpretation highlighting invariant syllables within these patterns? Our simulation study investigates how different strategies or cognitive processes map to task interpretations in more detail.

Until now, the algebraic phenomena have been researched in infants for a considerate amount of time (see Rabagliati et al., 2018). Similar to how grammatical rules like subject–verb–object structure are abstract and can apply to various specific words, algebraic learning is seen as reflecting a rudimentary exemplar for abstract syntactic processing. Recent meta-analyses have revealed an overall small effect of algebraic learning, which is more pronounced in meaningful contexts (Rabagliati et al., 2018). Furthermore, when meaningfulness is controlled for, there is a marginal age-related decline (Rabagliati et al., 2018). Despite the direct evidence on algebraic performance being primarily based on preferential focusing rather than cognitive processes, the belief that infants can acquire abstract rules has spurred decades of computational studies investigating the counter-argument of whether lexical learning can give rise to rule learning (Alhama & Zuidema, 2019). Recent simulations of algebraic phenomena have shown that the observed preferential focusing in experiments can sometimes be simulated by connectionist models (Alhama & Zuidema, 2018). Thus, whether young children learn abstract rules remains a matter of debate, and the underlying mechanisms of algebraic performance need to be clarified.

2.3 Recent theoretical perspectives

The aforementioned studies demonstrate the diverse strategies young children use across different linguistic phenomena. When learning words with unique syllables, young children appear to gradually progress from individual syllables to understanding word forms. In contrast, when learning simple syntactic structures, they may adopt either syntactic or lexical strategies depending on the task context. This aligns with Frank et al. (2016), who found that young children can flexibly adapt their processing sequences to the task environment at hand. This perspective challenges the overly rigid view that separates lexical learning and syntactic processing as orthogonal strategies.

Alternatively, recent research suggests that detecting syllable repetition may not be a high-level syntactic ability but rather a simpler perceptual strategy called *sameness detection*. This strategy allows young children to identify matches between the individual

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

syllable within a perceived syllable sequence (i.e., content currently held in auditory working memory) to the immediate perception of any syllable from the task environment (de la Cruz-Pavía & Gervain, 2021; Endress et al., 2009). Notably, this strategy emerges early in development. Evidence suggests that even newborns can detect simple repetitions and learn basic lexical forms, as supported by neural studies (Bouchon et al., 2015; Gervain et al., 2008) and comparative studies across species (for a review, see Wilson et al., 2018).

Beyond sameness detection, research suggests another strategy that allows young children to detect pattern-level mismatches. This strategy involves comparing expected patterns that were previously learned (i.e., long-term memory) with those currently encoded in auditory working memory (de la Cruz-Pavía & Gervain, 2021). While it is traditionally assumed that lexical learning precedes syntax, behavioural evidence shows a later developmental onset for lexical learning, typically around 6–9 months, compared to the earlier emergence of the sameness detection (de la Cruz-Pavía & Gervain, 2021). Neural evidence suggesting newborns lacking response to lexical patterns further corroborates the delayed development of lexical learning (Bouchon et al., 2015). Note that while our study focuses on basic strategies in simple tasks, acquiring more realistic syntactic abilities happens later. For instance, recognising complex non-adjacent dependencies like present continuous verbs ("is learning") typically emerges around 17 months (Gómez & Maye, 2005; Gómez, 2002), requiring both syntactic understanding and specific lexical forms.

While previous research has suggested that children use different strategies (e.g., Marcus et al., 1999; Saffran et al., 1996), the evidence supporting these claims has mainly come indirectly from behavioural findings, especially preferential focusing time. Behavioural preferential focusing is a dynamic process that changes over the course of learning. The Hunter-Ames model has summarised the learning-related directional biases of preferential focusing dynamics in relation to habituation levels (Hunter & Ames, 1988). When habituation is inadequate, such as with short training phases or complex tasks, children often show a familiarity preference. This means they tend to look longer at the familiarised, repeated pattern compared to a novel one. However, in most well-designed tasks, successful habituation is typically achieved through longer training or simpler tasks. This leads to a novelty preference, where children look longer at the novel pattern. Recently, the Hunter-Ames model has also been invoked to explain learning-related neural responses (e.g., Emberson et al., 2019; Issard & Gervain, 2018). As preferential focusing provides only an indirect indication of strategy acquisition, further research is needed to elucidate the relationship between them.

Can we then infer strategy learning based on the continuous trend of preferential dynamics? Alhama and Zuidema (2019) have a compelling review on this topic, positing that the behavioural results of early linguistic acquisition offer at least two potential interpretations. One perspective, shared by the majority of authors, holds that preferential focusing can directly underpin the successful acquisition of determinate strategies. Another possibility is that young children may have implicitly developed sufficient expectations regarding the training pattern, enabling them to differentiate between learned and novel patterns. However, this does not imply that they have fully grasped the anticipated strategies. Perhaps, partial ongoing procedural learning can also result in differences in preferential focusing. In this study, our computational simulations allow us to differentiate these two interpretations and determine whether preferential focusing truly reflects strategy acquisition or not.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

3. Unifying statistical learning and algebraic perspectives

Our brief review will describe various computational models for early linguistic acquisition. For each, we will highlight its core strategy or cognitive processing (e.g., chunking, syllable mismatch detection) and its related task interpretation (e.g., predicting adjacent syllables or multisyllabic patterns). We will not delve into specific implementation details (e.g., connectionist, syllabic, mathematical), which are better explored in technical references. This review aims to illustrate each model's specialisation in particular phenomena but also to point out the inherent trade-off – their design with specific processing assumptions (e.g., lexical patterns are chunked from syllables rather than parsed from continuous utterances) often limits their flexibility in adapting to different tasks or task changes.

Various computational approaches have been used to investigate the phenomenon of statistical learning. Following the transitional probability assumption, for instance, Christiansen et al. (1998) initially showed that a simple recurrent neural network (RNN) could predict the subsequent syllable based on its preceding context. The model is also capable of identifying word boundaries when the predictability of the next syllable decreases (e.g., parsing after syllable *Z* of *X-Y-Z* based on transitional probability). Nevertheless, the model's inherent mechanism precludes the acquisition of multisyllabic patterns (e.g., identifying and chunking the *X-Y-Z* triplet as a word, readily linkable to a referent), a limitation implied by the empirical findings of Estes et al. (2007).

Memory-based models aiming to acquire multisyllabic lexical patterns typically rely on (a) a parsing mechanism and (b) a mechanism for lexical learning and inference. For instance, the mathematical model of PARSER (Perruchet & Vinter, 1998) applies a parsing mechanism that randomly concatenates currently presented syllables (e.g., appending a 1–3 syllable window) and stores them in long-term memory. The connectionist TRACX models (French et al., 2011; Mareschal & French, 2017) enhance the parsing mechanism with a syllable mismatch detection system. Initially, individual syllables (e.g., *X*) are introduced to an input layer and subsequently transcribed to a memory layer. As an input sequence (e.g., *X-Y-Z*) is repeatedly encoded, an input syllable that is part of an established pattern (e.g., input *Z*) will not trigger a significant mismatch error, provided it follows a relevant lexical feature (e.g., a stored *X-Y* pattern) in the memory layer. This decrease in error enables the formation of a chunk from the current input syllable and the memory content (e.g., *X-Y-Z*). Conversely, a larger mismatch error between the current input and stored information (e.g., input *X'* with a stored *X-Y-Z* pattern) indicates a syllable mismatch and prompts parsing. Note that this parsing assumption is consistent with the difference strategy proposed by de la Cruz-Pavía & Gervain (2021).

Beyond parsing mechanisms that allow for memorising multisyllabic patterns, an inference mechanism is also needed to select the most appropriate pattern given the current context. In the PARSER model, the suitability of stored lexical patterns is inferred from the co-occurrence frequencies of concatenated forms and the features of currently presented syllables (Perruchet & Vinter, 1998). However, co-occurrence frequency alone can assign similar activation to equally plausible patterns from learning history, hindering the prediction of the most contextually relevant pattern (see Hoppe et al., 2022). Alternatively, TRACX models rely on an associative mechanism for lexical inference that accounts for predictions based on immediate input (see French et al., 2011; Mareschal & French, 2017). This makes the model's predictions more contextually relevant. Inference mechanisms can also incorporate a discriminative mechanism that better infers certain

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

patterns over others (see Hoppe et al., 2022). For instance, the iMinerva model (Thiessen & Pavlik, 2012) includes this additional assumption.

While the models discussed above can simulate lexical learning at syllable and/or multisyllabic levels, they have limitations. One such limitation is the assumption that lexical learning originates from small units like syllables. This prevents them from explaining alternative empirical findings, such as instances where young children parse lexical patterns from continuous utterances without focusing on individual syllables (see Arnon, 2021). Furthermore, early linguistic development involves more than just lexical learning; it also includes phenomena like algebraic learning. To address these distinct phenomena, a model needs to incorporate different strategies. One modeling approach, exemplified by the symbolic ideal observer approach (Frank & Tenenbaum, 2011), predefines a hypothesis space of strategies as a collection of rules. For instance, a chunking rule for *X-Y-Z* could be simply represented as <is_*X*,is_*Y*,is_*Z*>. Another rule, <_, _, =1>, indicates that the last syllable is identical to the first, and can process the algebraic pattern *a-b-a* where the first and third syllables match. The appropriateness of rules for a given task is also inferred statistically. This particular model uses a Bayesian approach for rule inference, predicting the current task based on updated rule applications. In other words, increased successful rule application leads the model to anticipate and apply that rule to process a given syllable sequence.

While distinct rules may correspond to different strategies underpinning linguistic phenomena, they have several disadvantages, including inflexibility and uncertain cognitive plausibility. Regarding inflexibility, rules are fixed and therefore cannot model the gradual exploration from partially inappropriate rules to those more suitable for the current task scenario. Additionally, some patterns, depending on how they are processed (e.g., syntactic *a-b-a* or lexical *a-X-a*), can lead to multiple interpretations, would potentially reduce a certain rule's robustness. It is also not straightforward how a rule like <_, _, =1 > would map onto plausible cognitive processing in infants. Last but not least, for some of the rules in the hypothesis space, lexical context must be integrated into the rules (e.g., <is_*X*>), making them context-dependent rather than general purpose. To better understand rule learning and address the preference for a model to learn the hypothesis space of rules from scratch, a hybrid model incorporating cognitively plausible processing and statistical learning is perhaps needed (Alhama & Zuidema, 2019; Frank et al., 2016).

In this article, we introduce the PRIMs model, which achieves cognitive plausibility by adopting a general cognitive architecture with various modules (input, working memory, declarative). Furthermore, PRIMs incorporates a set of general-purpose processing elements (primitive operators) for information movement and comparison within this architecture. These operators initially tackle a given task openly through trial and error, producing a resulting sequence of processing steps termed a *procedure* (described shortly). By interacting with the task at hand, the model can develop processing sequences that resemble the lexical chunking mechanisms of memory-based models (e.g., French et al., 2011; Mareschal & French, 2017; Thiessen & Pavlik, 2012) and also simple syntactic processing of rule-based models (e.g., French et al., 2011).

While the open discovery of processing steps relies on the processing assumptions of the cognitive architecture, the acquired *procedural representations* are statistical constructs that guide future task performance. These representations are conceptualised as a set of context-operator associations, which help to indicate which operator the model should select in the immediate context (described shortly) instead of anticipating an entire procedure at once (compare Frank & Tenenbaum, 2011). Applying an operator

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

updates the context, leading to the selection of the next operator, thus forming an operator sequence step by step. When a procedure is repeated, the collective set of contextual associations implicitly represents how the task should be processed.[1]

An additional feature of PRIMs is that it models the passage of time and can predict how long particular cognitive operations take. For example, it can predict when the speed of its processing cannot keep up with the speed at which it perceives speech. Contextual associations not only inform the selection of operators but also influence their efficiency of application. The processing rate is relative to the presentation rate of the task. Therefore, models with different efficiency levels may either be able to handle the task or result in partial processing due to omissions or inadequate procedures. This enables us to capture behavioural focusing dynamics from a resource availability perspective (Taatgen et al., 2021, described shortly).

3.1 Simulating early linguistic acquisition

Language acquisition can be viewed as the interplay between general cognitive abilities and the specific environment in which speech can be perceived. The architecture consists of a collection of sensory and central modules that process information relatively independently and communicate through buffers. The collective content of all the buffers determines the current context. The learning problem consists of discovering the right procedure in terms of a sequence of operators to move information between the buffers based on the context. For the purpose of this article, the modules depicted in Figure 1 are relevant. The sensory modules make sensory input available to the central workspace. Here we only focus on auditory input. The working memory module can be used to temporarily store information that is needed later in the process. This module is needed to detect repetition patterns in the input. The declarative module serves as long-term memory and stores previously perceived patterns. If partial patterns are placed in its buffer, it attempts to retrieve the most probable patterns that complete the partial patterns based on past experiences at the declarative retrieval buffer. This module is able to account for aspects of statistical learning. The goal module represents the current background environment. For the purposes of this study, it can be considered the setting of the experiment, but it can also represent a specific task that the model has set (described shortly in detail). Finally, there is an operator module that determines the flow of information through the central workspace. The directions of information processing are determined by operators, selected by the operator module, within the central workspace that connects all the buffers (see Figure 1). While not directly relevant to the current study, the model can also process information within the buffers and translate it into actions, such as vocalising the perceived or retrieved sequence of syllables.

When the cognitive architecture processes a novel speech stream, there is initially no content that can be retrieved to the declarative retrieval buffer, and there is also no content available in various buffers to be compared. During such a situation, the model may

[1]Note that a single set of context-operator associations, corresponding to a single procedural representation, can be easily leveraged to form skills involving multiple procedural representations and their (skill-skill) interconnections (e.g., in complex grammatical processing). However, for simplicity, we focus on a single set of context-operator associations in this study. This is because the targeted lab-based studies involve only a single learning phase that can be sufficiently modelled with the acquisition of simple procedures, rather than complex multi-procedure skills.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

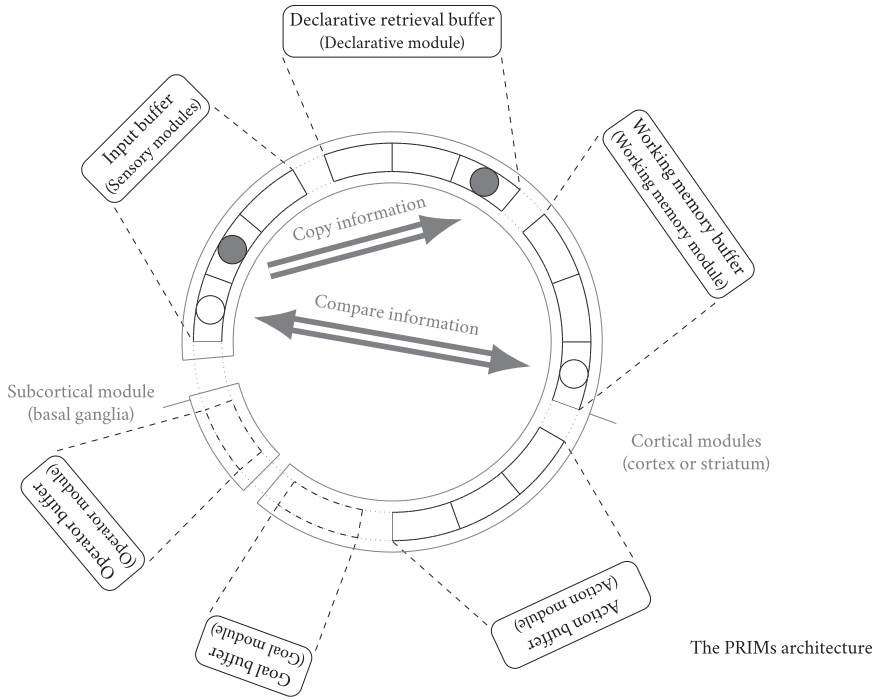


Figure 1. The PRIMs architecture. The PRIMs architecture, following its predecessors (e.g., SOAR and ACT-R), assumes that cognition can be decomposed into a set of specialised modules that are connected through a central workspace (sometimes called the global workspace, see Anderson, 2007; Dehaene et al., 1998; Taatgen, 2013). Each module projects onto a so-called buffer to communicate information via the central system to other modules. Within this structure, the primitive operators involve either comparing the available contents between two buffer slots (indicated by a double arrow) or encoding contents from one buffer slot to another empty buffer slot (indicated by a single arrow).

encounter retrieval failure and cannot issue comparison operators. Instead, it may gradually learn to encode the automatically perceived input content consecutively into the working memory buffer. The flexibly encoded sequence can then be stored in long-term memory, for instance, when the model encounters utterance boundaries (or inter-stimulus intervals). The stored content can then be made available in the declarative retrieval buffer through memory retrieval. The automatic placement of stimuli into the input buffer and the flexible application of encoding and retrieval operations allow information content to be available in various buffers, allowing comparison of buffer contents (e.g., input-working memory, input-declarative memory, and working memory-declarative memory comparisons). Based on previous literature (de la Cruz-Pavía & Gervain, 2021), we defined an input-working memory or input-declarative match as one that indicates sameness detection, while a working memory-declarative mismatch indicates difference detection. The model can freely choose any operator sequence that leads to these comparisons, and the final sameness or difference comparisons suggest that the model has recognised the presented pattern in some way.

Additionally, PRIMs differs from conventional cognitive architectures in how it handles operators. In conventional architectures, the operator triggered by a specific

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

"if" statement is stored in a separate system called procedural memory. In contrast, PRIMs utilises a general-purpose operator applicable in various scenarios, requiring only minimal conditions (e.g., the lack of buffer content for encoding and availability of buffer content for comparison). The selection of operators relies on associations with the current context. The context consists of the information in all the buffers, including the context of the experiment itself, which is assumed to be represented in the goal buffer. Immediate contexts are gradually associated with the corresponding operator when the model recognises the pattern based on sameness-based input-working memory or input-declarative matches or difference-based working memory-declarative mismatches. These learned associations gradually allow the model to select the right operator in immediate future scenarios to recognise the pattern.

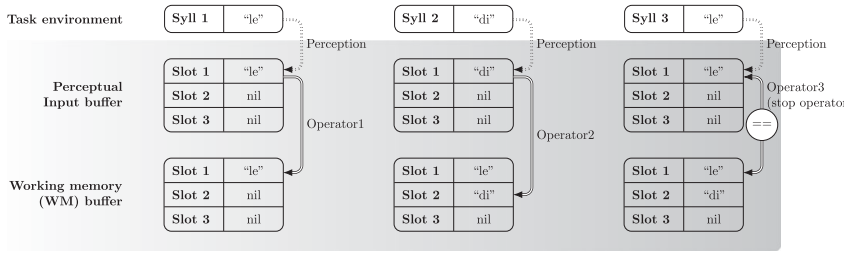
To illustrate the evolution of PRIMs' context-based procedural learning from rule-based approaches, we present concrete examples of procedural learning in an algebraic pattern (see Figures 2 and 3).

3.2 Procedural learning in algebraic task

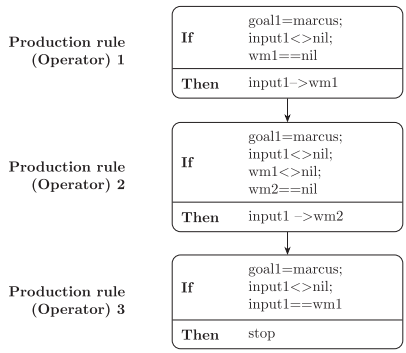
Let us look more closely at the information processing involved in an algebraic pattern, as illustrated in Figure 2a. The figure depicts the steps involved in processing the three presented syllables. In a conventional cognitive architecture, the processing sequence is determined by the modeller through production rules, as exemplified in Figure 2b. While production rules fulfil a similar role to operators, they achieve this by explicitly defining task-specific conditions and corresponding actions through if-then pairings. This example instead includes production rules with minimal conditions (only containing the specific condition, goal1 = marcus), which makes them similar to general-purpose operators. However, the approach remains rule-based, as the production rules cannot be learned to associate with their corresponding context. Instead, their suitability is informed by their utility frequency (Anderson, 2007). Furthermore, an operator sequence can be compiled into a single specialised task rule (Taatgen & Anderson, 2002), as shown by the solid arrows in Figure 2b.

Conversely, PRIMs employs a contextual learning mechanism that enables the model to flexibly acquire operator sequences. Unlike conventional cognitive architectures, PRIMs starts with a full set of general-purpose primitive operators that are not tied to specific conditions (see Supplementary Appendix for details). These operators just move one piece of information content from one buffer to another. The learning process involves identifying the appropriate operator for a given model state or context. Initially, operators are chosen through trial-and-error exploration. However, upon successful sequence completion (identified by specific stopping operators), PRIMs updates the association between each operator and the relevant context within that sequence (see Figure 2c). This context can encompass various aspects, including both "lexical" (i.e., current buffer content) and "syntactic" information (i.e., preceding operator), as the architecture remains neutral to the specific nature of the context. In subsequent encounters, the model leverages the learned contextual associations and the current model state or context to select the next operator. PRIMs focuses solely on whether the current context, regardless of its processing or lexical nature, is informative for guiding operator selection. Besides, in PRIMs, primitive operators can be merged into more complex ones, capable of performing multiple operations simultaneously. However, the process of automatically compiling adjacent operators is not considered in this article.

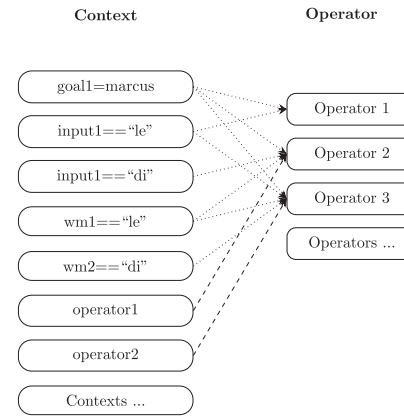
https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press



Panel A: A schematic diagram illustrates a possible operator sequence for a syllable repetition task within a cognitive architecture. The model hears "le" (perception) and stores it in working memory (encoding operator 1). It then perceives "di" (perception) and places it in the next slot (encoding operator 2). Finally, upon hearing "le" again (perception), it compares it to the first working memory slot (comparison operator 3). This content match signifies successful task completion (repetition procedure).



Panel B: Conventional cognitive architecture relies on production rules. These rules establish rigid if-then relationships to guide task processing. Each rule is essentially a pre-programmed pairing of conditions and actions, determined by the modeler beforehand. A key limitation of this approach is the lack of a clear mechanism for the model to learn these crucial if-then pairings.



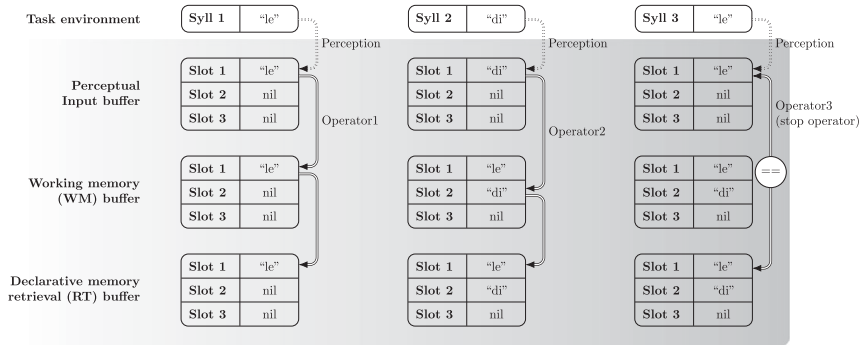
Panel C: PRIMs applies a flexible set of general-purpose operators. Once an operator sequence is complete, and signaled by a stopping operator, the associations between the operators and the encountered contexts are strengthened (see dotted/dashed arrows). Here, the operator context (e.g., operator1) carries more weight than the specific, variable lexical content (e.g., input1="le").

Figure 2. The processing of an algebraic pattern. The upper panel provides an overview of the processing steps within a cognitive architecture. The lower panel compares traditional cognitive architecture with PRIMs.

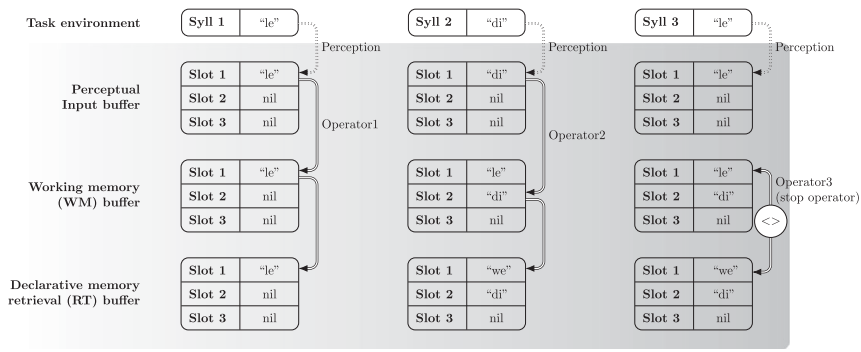
Instead, the extent to which operators are merged is determined by the strength of the operator–operator association. In the next section, we will also look at how contextual learning within a processing framework sheds light on preferential focusing dynamics.

While pre-defined rules might achieve the sequence described earlier, PRIMs is also capable of learning alternative plausible operator sequences. Initially, the model lacks stored declarative items to be retrieved. It may then encode input into working memory

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press



Panel A: In this example, when information is encoded from the input and placed into working memory (encoding operators 1 and 2 in Figure 2), it also triggers memory retrieval. This retrieval request involves placing the contents into the declarative retrieval buffer, as indicated by the additional arrows from working memory to declarative memory. The retrieved content is then placed into the retrieval buffer (like "le" or "le-di" based on the first two encoding operators). Based on the last comparison operator, the model identifies a position-independent match between the perceived input and the retrieved content from long-term memory (input1==rt1). This comparison differs from the input-working memory comparison (input1==wm1) seen in Figure 2, but both of these operator sequences belong to sameness detection (repetition procedure).



Panel B: In this alternative example, memory retrieval is also involved in retrieving stored content in the declarative retrieval buffer. Based on the last comparison operator, the model identifies a mismatch between the encoded working memory content and the retrieved content from long-term memory after the second syllable (i.e., learning 1-gram), specifically at the first slot position (wm1<>rt1). This position-specific mismatch indicates a pattern-level deviation between the currently perceived sequence and expectations from experience. The operator sequence thus belongs to difference detection (n-gram procedure). While difference detection is presented as an alternative procedure in this algebraic example, it is essential for learning statistical learning patterns where the order of individual lexical contents is crucial.

Figure 3. Alternative processing steps for the same algebraic pattern. *Note*: the current model assumes that working memory encoding occurs spontaneously with long-term memory retrieval. Therefore, exogenous working memory encoding operators 1, 2, and 3 are automatically followed by a retrieval request that endogenously encodes content from long-term storage into the retrieval buffer slots.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

and compare input and working memory content to recognise patterns. However, as learning progresses, the model gradually stores lexical forms and leverages past experiences to infer patterns. In PRIMs, the declarative retrieval buffer facilitates lexical inference by retrieving previously stored information from long-term memory when presented with partial patterns. The enhanced capability in lexical inference introduces another repetition procedure (sameness detection) illustrated in Figure 3a. Likewise, this repetition procedure focuses on individual syllable features, disregarding their specific order within the sequence. Consequently, processing patterns like "le-le," "le-di-le," or "le-di-we-le" (all ending with "le") trigger similar match operators, showing the limitation of this procedure in capturing the overall lexical structure of the pattern.

Crucially, pattern inference further empowers the model to consider the perceived sequences as lexical forms. This enables a direct comparison between the perceived syllable sequence in working memory and the retrieved sequence from long-term memory. This difference detection typically encounters mismatches at lexical boundaries, reflecting the uncertainty in inferring the next word after a fixed one. However, the specific location of the mismatch is irrelevant. As illustrated in Figure 3b, upon reaching a specific encoded sequence position, any mismatch (e.g., wm1<>rt1) indicates a difference at the pattern level between the perceived and retrieved sequences. This study investigates how the detection of these position-specific differences relates to the formation of corresponding *n*-grams. In this example, identifying a mismatch after perceiving the second syllable creates a 1-gram. Similarly, detecting mismatches after perceiving the third, fourth, and subsequent syllables corresponds to 2-gram, 3-gram, and *n*-gram procedures, respectively. Moreover, details regarding the primitive encoding and comparison operators, as well as the predefined comparison operators that define the stopping conditions, can be found in Supplementary Appendix.

3.3 Preferential focusing dynamics

We have previously mentioned that PRIMs utilises a contextual learning mechanism to update the associations between contexts and the operators they trigger. In future situations, the model uses these learned associations and the current contextual state of the architecture to predict the most suitable operator to perform. This gradual strengthening of contextual associations allows the model to become more familiar with the task environment by strengthening suitable context-operator associations. The following will provide an alternative processing-based explanation (for discussions, see Houston-Price & Nakai, 2004) for the U-shaped trajectory of preferential focusing dynamics (Hunter & Ames, 1988) based on a resource availability perspective (Taatgen et al., 2021). A further implication of the processing-based perspective is related to how we interpret strategy acquisition based on preferential dynamics. From a contextual learning perspective, ongoing changes in contextual associations influence the efficiency of task processing, leading to differences in preferential focusing. Therefore, preferential focusing differences do not necessarily indicate a binary state of strategy acquisition (either acquired or not; see Alhama & Zuidema, 2019).

Processing-based interpretations. In familiar and repetitive task environments, the model develops stronger expectations about the task by learning which operators to apply in specific contexts. This learning process strengthens the contextual associations corresponding to repeatedly successful operators. A stronger context-operator association leads to a higher total association, which ultimately influences the activation level

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

(likelihood of selection) of the corresponding operator. Operator activation, in turn, directly affects the time taken to select an operator (see Supplementary Appendix). As the model becomes increasingly familiar with the task environment, a gradual shift occurs from slower processing (due to lower activation) to faster processing (due to higher activation). This increased processing speed enables the model to apply more operators per unit of time, potentially allowing it to transition from frequent omissions or partial procedures to more complex procedures that recognise the overall structure of the task. This increasing allocation of task processing time may resemble the phenomenon of familiarity preference, where attention is increasingly directed towards the task at hand (Hunter & Ames, 1988).

However, once the model has achieved proficiency in processing the current task, it may enter brief periods of inactivity (e.g., after encoding a syllable before the presentation of a subsequent syllable). During these idle times, the model can potentially explore new information or engage in processes unrelated to the immediate task. This shift away from the primary task is similar to the observed decrease in preferential focusing towards familiar stimuli and increase in preferential focusing towards novel stimuli (Hunter & Ames, 1988). After sufficient training/habituation, familiar tasks benefit from faster processing, resulting in reduced focusing times. However, encountering novel tasks in test conditions may again alter operator activation, impacting immediate operator efficiency. For instance, a small change (e.g., syllable change) in a learned sequence disrupts the established context, reducing activation and efficiency. Similarly, new tasks requiring unfamiliar operators lack strong contextual associations, again affecting activation and efficiency. This can then explain preferential focusing under various task conditions.

In our model, on-task processing time approximates focusing time, defined as the total time spent on essential operators within a procedure. For example, in Figure 2, this would involve only the integral operators 1, 2, and 3, excluding other operators that may be applied flexibly. Efficient enhancement and reduced on-task processing time are closely linked, which may lead to a further decline in overall focusing time due to competing off-task processing (e.g., mind wandering or following alternative operator sequences, see Taatgen et al., 2021). Nevertheless, disengagement can readily occur whenever operator-level temporal resources allow for an attentional shift. At a finer level, preferential focusing dynamics is thus related to moment-by-moment operator latency, which reflects real-time changes in temporal resources and the tendency to disengage. Note that for convenience, the current study considers only on-task processing and does not account for attentional competition, such as attentional disengagement or reorientation to off-task or novel activities.

Developmental factors. Developmental factors also influence focusing dynamics. Due to myelination, older children process more efficiently than younger children (see Dubois et al., 2014). For example, Chen et al. (2016) demonstrated that younger children may omit or incompletely process the middle tone when presented with a simple three-tone sequence. Moreover, while meta-analyses do not show an age-related trend in the effect sizes of preferential focusing differences (Isbilen & Christiansen, 2022; Rabagliati et al., 2018), developmental evidence indicates that the speed of habituation (i.e., the decrease in focusing across training trials) is faster in older infants compared to younger infants (Dawson & Gerken, 2009; Frank et al., 2020).

Task-specific age-related trends have also been revealed. For statistical learning, preferential focusing is positively correlated with age during infancy (Emberson et al., 2019) but diminishes among older children or adults (Frost et al., 2019; Isbilen & Christiansen, 2022). In contrast, for algebraic processing, preferential focusing is reduced

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

among older children compared to younger children (Dawson & Gerken, 2009). Although meta-analyses suggest such an effect is only marginal after controlling for meaningfulness (Rabagliati et al., 2018).

In PRIMs, the differential age-related findings can be simulated by incorporating models with fixed levels of efficiency. Note that we acknowledge that developmental maturation is co-determined by learning (e.g., see Huber et al., 2023) but assume that such structural changes occur at a much slower rate than learning at a functional level. When efficiency is low, the model may not be able to retrieve patterns from long-term memory (declarative retrieval). This suppresses statistical learning, where long-term memory is crucial (Figure 3b), but benefits algebraic processing, where having working memory content available is sufficient (Figure 2). An extremely slow model can lead to syllable omission, further hindering lexical learning. However, even with syllable omission, repetition detection remains possible, for instance, when the middle token *b* in an *a-b-a* pattern is omitted.

3.4. Research aim and questions

Simulation studies 1 and 2 first examine the lab-based statistical learning (Saffran et al., 1996) and algebraic (Marcus et al., 1999) paradigms, along with their contrasting age-related trends (Emberson et al., 2019; Isbilen & Christiansen, 2022; Rabagliati et al., 2018). The study focuses on processing efficiency, moderated by parameters related to age-related development and the learning experience (described shortly). In these studies, we consider the question of whether preferential focusing dynamics provide evidence for all-or-none explicit strategy acquisition or the continuous implicit learning of context-based procedural representations (see Alhama & Zuidema, 2019). Beyond the simulation of lab-based paradigms, we leverage the model to explore more naturalistic word learning (or the acquisition of word-level phonological patterns) by exposing it to infant-directed speech. We also simulate individual differences in word learning trajectories by moderating the model's efficiency levels.

4. Simulation results and discussion

The technical details of the PRIMs architecture and the key model parameters of this study can be found in Supplementary Appendix. Please also refer to https://git.lwp.rug.nl/y.ji/prims-contextual-learning for the software, model scripts, and analysis codes. Studies 1 and 2 focus on the aforementioned lab-based paradigms, while Study 3 focuses on leveraging the model to simulate more naturalistic word learning from infant-directed speech settings. For the simulation studies, the section begins with a description of the task design and conditions.

In Studies 1 and 2, we further address the question raised by Alhama and Zuidema (2019) of whether preferential focusing dynamics implies sufficient strategy acquisition. Therefore, we begin with simulation results related to preferential focusing, such as processing time and operator latency, to determine if they are consistent with the observed empirical findings. This is followed by additional analyses of the procedures learned and their corresponding proportions of application. These assess whether the results support complete strategy acquisition or only partial learning and differentiation of procedures. In Study 3, we do not consider preferential focusing and procedural learning but instead focus directly on the word learning outcomes of the model. This

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

involves examining which word-level phonological patterns have been acquired and how the average activation trajectories of these patterns change over the learning phases.

Operator latency in all simulations is influenced by both developmental and learning-based components. The *developmental component* reflects structural changes mediated by myelination or brain maturation (e.g., see Dubois et al., 2014). Brain maturation, while co-influenced by learning, has a slower rate than immediate learning experience. Our current model simulates developmental change by manipulating the *action time*, which is the time it takes to apply an operator. As depicted in Figure 1, an operator is the processing element that moves or compares information content between buffers, representing long-range connections between cortical areas based on the striatocortical system. By default, the action time is fixed to 50 ms, but in our simulations, we increased the action time to represent younger infants. The *learning-based component* reflects that with practice, the operator sequences become more efficient and take less time. To isolate learning-based efficiency from other factors, the model eliminates additional fixed base-level activation, explore–exploit scaling of activation noise, prim- and operator-level compilation, along with declarative retrieval latency (for more information, see PRIMs tutorial in Taatgen, 2023). Since context-based operator latency is transformed exclusively from activation (i.e., A) in the current model (based on $F \cdot e^{-A}$; see Equation (A7), Supplementary Appendix). Thus, operator efficiency is higher when operator activation (i.e., total learned contextual association of an operator) is higher. In all simulations, the latency factor scalar F is set to 200 ms, resulting in an initial context-based latency of 200 ms for all operators across model runs ($200 \cdot e^{0} = 200\,\text{ms}$).

4.1 Study 1: Lexical statistical learning

This simulated task is adapted from Saffran et al. (1996). The training involves exposing the model to a set of fixed-token trisyllabic words following the *X-Y-Z* format. During testing, the model has to distinguish these learned words from novel words constructed using the same set of tokens.

Task material. The training phase consists of four trisyllabic words (e.g., "pa-bi-ku," "ti-bu-do," "da-ro-pi," and "go-la-tu"). These words are randomly linked together to form a continuous stream of syllables without intervals between them. After the presentation of continuous syllables, the trained model is each followed by a word, non-word, and part-word test phase.[2] The consistent word test condition selects two test words from the training phase (i.e., "pa-bi-ku" and "ti-bu-do"). From the perspective of transitional probability, as we mentioned earlier, the word condition fully converges with the transitional probability of the training phase ($p = 1$). The part-word condition combines the last syllable of one word with the first two syllables of another, reflecting word boundaries (e.g., "tu-da-ro" and "pi-go-la"). For an illustration of part-words, take the phrase "pretty baby." A part-word would be a syllable sequence like "tybaby," which spans

[2]Note that the original experiment applies a within-subject design when it comes to different test conditions. This, beyond reducing between-subject variability, measures how focusing time differs between conditions. This design would further require the order of the test conditions to be counterbalanced. In the simulation study, however, a between-subject design is applied, where each training phase is followed by only a single test condition. The approximate focusing preference is based on the differences in processing time and is further informed by the moment-by-moment operation latency, as described earlier. Therefore, the simulation study does not account for order effects or potential competition between different test conditions, which might further influence preferential focusing.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

the word boundary between "pretty" and "baby." The part-words partially converge with the transitional probability of the training phase ($p = 1/3$). The non-word test condition still contains the syllable tokens of the training words (e.g., "da-pi-ku," "ti-la-do"), but the positional information of the syllable tokens within the training words is completely disrupted, making the non-word condition completely differ from the transitional probability of the training phase ($p = 0$).

Presentation and processing rate. Each trisyllabic pattern is presented for 1500 ms, resulting in a syllable presentation rate of 500 ms per syllable without any intervals between patterns.[3] The model includes two efficiency levels to model developmental trends. The low-efficiency level allows the model to apply only a single operator during the presentation of a single syllable (i.e., action time: 300 ms; latency factor: 200 ms). In contrast, the high-efficiency level allows the model to choose more operators during the unit time of a syllabic presentation (i.e., action time: 50 ms; latency factor: 200 ms). The models with different efficiency levels represent processing rates that are comparable to younger and older infants (for more information, see Chen et al., 2016; de la Cruz-Pavía & Gervain, 2021; Di Liberto et al., 2023). Each model run incorporated 200 training patterns and 10 test patterns. The simulation study included 100 separate model runs for each of the word, non-word, and part-word conditions.

Simulated preferential focusing. The dependent variables were examined at two levels. Overall *on-task processing time* provided an approximation of preferential focusing time, calculated as the sum of latencies from all essential operators throughout the entire test phase. The second dependent variable was averaged *operator-level latency*, which captures moment-by-moment changes in temporal resources relating to attentional disengagement. Crucially, within our analyses, we focused only on the context-based component of operator latency. This specific component is particularly important because, unlike the fixed action time, it directly reflects the model's learning mechanisms and its dynamic, context-driven adjustments. In this study, we analyse two independent variables: *test condition* (word, part-word, or non-word) and *model efficiency* (high or low), which represents age-related development (younger versus older infants) and was operationalised as high efficiency (low action time) versus low efficiency (high action time).

Linear regression analyses were performed in R (version 4.0.2) to examine the main effects of test condition and model efficiency, and their interaction, on both the on-task processing time and the context-based component of averaged operator latency. Our approach for all analyses involved comparing four nested linear models: Model 1 and 2, respectively, incorporated the main effects of test condition and model efficiency; Model 3 comprised both main effects; and Model 4 additionally included the interaction effect. A forward-feeding model comparison approach was utilised to analyze whether the inclusion of a specific main effect (e.g., model efficiency in Model 3, while controlling for test condition as in Model 1) enhanced model fit. This same approach was subsequently applied to evaluate the contribution of the interaction (e.g., Model 4 in addition to Model 3's main effects).

[3]Note that the presentation duration of syllables in the simulation is slightly longer than that in the original experiment (i.e., 300 ms). This is to reduce the likelihood of the model ignoring syllables and resulting in insufficient processing (see Chen et al., 2016). Furthermore, the trisyllabic patterns in the test phase of Saffran et al. (1996) were separated by intervals. However, in our simulation experiment, we consistently adopt continuous syllable presentation in both the training and test phases. The reason for this is that our model depends on a potentially mismatched fourth syllable (i.e., X' in X-Y-Z-X'-Y') to identify the boundary of the trisyllabic 3-grams.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

On-task processing time. A significant improvement in model fit is found when considering the main effects of test condition (compare Models 2 and 3; $F_{(2,596)} = 23.17$, $p < 0.001$, $\Delta AIC = -40.93$) and model efficiency (compare Models 1 and 3; $F_{(1,596)} = 1886.20$, $p < 0.001$, $\Delta AIC = -854.00$). Model fit is further enhanced when the interaction effect is added (compare Models 3 and 4; $F_{(2,594)} = 8.46$, $p < 0.001$, $\Delta AIC = -12.85$). Therefore, we decide that Model 4 is the best-fitting model.

Model 4 reveals a significant difference in on-task processing times between word, part-word, and non-word conditions in high-efficiency simulations ($\beta_{\text{w.-p.w.}} = -0.44$, $SE = 0.12$, $t = -3.58$, $p < 0.001$; $\beta_{\text{n.w.-p.w.}} = 0.52$, $SE = 0.12$, $t = 4.17$, $p < 0.001$), a distinction not revealed in low-efficiency simulations. This means, for high-efficiency simulations, on-task processing time is 442 ms faster for words compared to part-words, and 515 ms faster for part-words compared to non-words. On-task processing times are 3188 ms shorter overall for high-efficiency simulations than for low-efficiency simulations ($\beta_{\text{low-high}} = 3.18$, $SE = 0.12$, $t = 25.84$, $p < 0.001$). Furthermore, the interaction coefficients suggest that the differences between conditions in the high efficiency conditions are reduced in the low efficiency condition ($\beta_{\text{[low-high]:[n.w.-p.w.]}} = -0.44$, $SE = 0.17$, $t = -2.52$, $p = 0.01$; $\beta_{\text{[low-high]:[w.-p.w.]}} = 0.27$, $SE = 0.17$, $t = 1.56$, $p > 0.1$). This means that when the model is less efficient, the difference between non-word and part-word conditions is 440 ms smaller than when the model is more efficient, basically reducing the difference found in the high efficient simulation (515 ms). Similarly, the difference between words and part-words, which was 442 ms in the high efficient simulations, is reduced with 271 ms in the low efficient simulation. Note that the interaction coefficients are associated with larger standard errors and therefore may not found to be significantly different from zero. Nevertheless, the model predictions clearly show a strong interaction effect: The test conditions are only resulting in different processing times in the high efficiency simulations (95% CI for w.: [2290, 2630], p.w.: [2730, 3070], n.w.: [3250, 3590])., but not in the low efficiency simulations (95% CI for w.: [5750, 6090], p.w.: [5920, 6260], n.w.: [6000, 6340]).

Context-based operator latency. The model fit is significantly improved by incorporating the main effects of test condition (compare Models 2 and 3; $F_{(2,596)} = 9.17$, $p < 0.001$, $\Delta AIC = -14.18$) and model efficiency (compare Models 1 and 3; $F_{(1,596)} = 706.99$, $p < 0.001$, $\Delta AIC = -467.30$). However, adding the interaction effect did not increase the model fit (compare Models 3 and 4; $F_{(2,594)} = 1.75$, $p = 0.17$, $\Delta AIC = 0.47$). Therefore, we decided that Model 3 is explaining the context-based operator latency best.

Model 3 reveals that in low and high-efficiency simulations, context-based operator latencies differ significantly between word and part-word conditions ($\beta_{\text{w.-p.w.}} = -0.0013$, $SE = 0.0004$, $t = -3.18$, $p = 0.002$), with word latencies being 1.3 ms faster. No significant difference was found between non-word and part-word conditions in the low and high-efficiency simulations. However, low-efficiency simulations show larger overall latencies ($\beta_{\text{low-high}} = 0.0090$, $SE = 0.0006$, $t = 26.59$, $p < 0.001$), indicating that averaged operator latency was 9.0 ms faster in high-efficiency compared to low-efficiency simulations.

Summary. In this study, on-task processing time serves as an indicator of preferential focusing, while operator latency reflects the moment-by-moment tendency of engagement. In addition, the efficiency level in our study corresponds to an age-related factor. Our simulation results show that in the high-efficiency simulations, on-task processing time is fastest for words, followed by part-words and then non-words. Similarly, context-based operator latency is more efficient for words than for part-words and non-words (see Figure 4). These simulated results align with novelty preferences observed in empirical

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

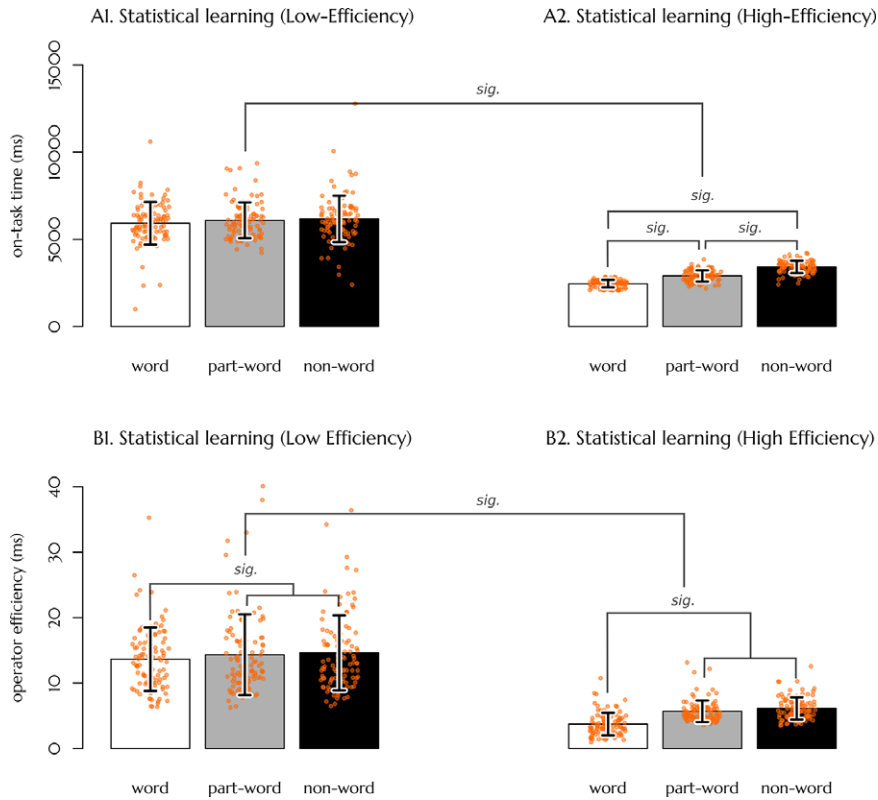


Figure 4. Simulated preferential focusing dynamics regarding statistical learning tasks. (a) Averaged processing time under test conditions. Calculated as the sum of on-task operator latencies during the test phase. *Note.* y-axis: average processing time (in ms); x-axis: test conditions; white bars: consistent conditions; various gray bars: inconsistent conditions; error bars: ± 1SD; orange dots: data points from individual model runs. (b) Averaged operator efficiency under test conditions. Calculated based on the average context-based latency (excluding fixed default action time) of all operators across model runs. *Note.* y-axis: average latency (in ms); x-axis: test conditions; white bars: consistent conditions; various gray bars: inconsistent conditions; error bars: standard deviations; orange dots: data points from individual model runs. The significance of regression coefficients is denoted by brackets and indicators (sig., $p < 0.001$). Note that the overarching brackets denote the main effect of model efficiency.

literature (e.g., Emberson et al., 2019; Estes et al., 2007; Saffran et al., 1996) and indicate that statistical learning effects typically emerge in older rather than younger infants (see de la Cruz-Pavía & Gervain, 2021). In addition, higher efficiency leads to shorter on-task processing time and operator latency compared to low-efficiency simulations. The simulated results are thus consistent with the main effect of age on preferential focusing time in general (e.g., Dawson & Gerken, 2009; Frank et al., 2020).

In the simulation study, the differences in on-task processing time and operator latency between test conditions correspond to preferential focusing bias and immediate engagement differences, respectively. The age-related levels of model efficiency only moderately simulated preferential bias (the between-condition on-task processing time difference). This is consistent with empirical findings that suggest an age-related enhancement of preferential focusing in statistical learning (e.g., see Emberson et al.,

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

2019). However, model efficiency does not influence the difference in engagement (the between-condition operator latency difference), which corresponds to meta-analytical findings that cast doubt on such developmental effects (see Isbilen & Christiansen, 2022).

Underlying procedural learning. Figure 5a and b illustrates the performance of the low- and high-efficiency models. Figure 5a/b1 depicts the proportion of procedures applied across learning blocks during the learning phase, while Figure 5a/b2 depicts the proportion of procedures applied in a single block under various test conditions following the training phase. Note that Figure 5a1/b1 depicts the training phase preceding the word test condition only.

Training phase. The low-efficiency model fails to acquire the appropriate 3-gram procedure (pale triangle, reaching 0.2%, see A1), while the high-efficiency model favours the 3-gram procedure at the end of training (reaching 45.2%) along with 1-gram (pink hourglass) and 2-gram (brown diamond) procedures applied during the initial training phase (see Figure b1). The high-efficiency model's initial transitions from 1-gram to 2-gram procedures correspond to learning from single syllables to predicting the immediate next syllable (transitional probability). Once the trisyllabic pattern can be correctly inferred, the model consistently applies the 3-gram procedure (phonological form).

The differential efficiency-related procedural learning results can be interpreted from a resource availability perspective (Taatgen et al., 2021). When efficiency is low, the model

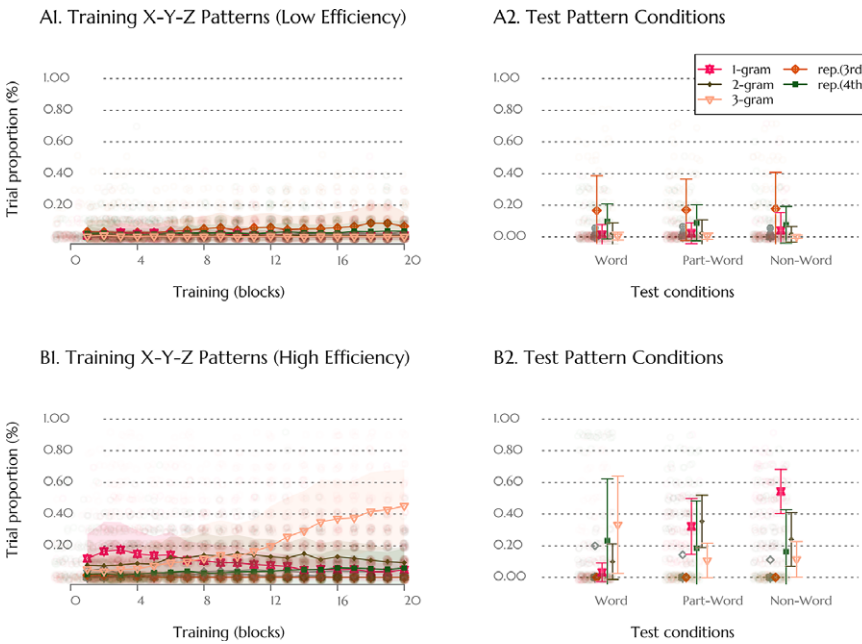


Figure 5. The averaged proportion of procedures applied by the model in the statistical learning task across model runs. The *n*-gram procedures detect differences between the working memory and the retrieved pattern at the *n* + 1th position. The first repetition procedures detect repetition of input syllables compared to an already encoded working memory slot (slot 1) at the 3rd (orange diamond) and 4th (green square) syllable positions, respectively. *Note.* y-axis: averaged trial proportion of each procedure (within each block); training-phase x-axis (A1/B1): 20 blocks each consisting of 10 trisyllabic patterns; test-phase x-axis (A2/B2): test conditions; error band/bars: ±1SD; transparent dots: data points from individual model runs. The sum of the trial proportions is not equal to 1, as the model may not use any procedure or may use more than one procedure in a trial.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

can only use one operator for encoding and lacks the temporal resources to make the necessary comparisons for *n*-gram procedures. In extreme cases, syllables might be completely ignored without being encoded. Conversely, the high-efficiency model allows the application of multiple operators in a single-syllable presentation window, which supports the application of *n*-gram procedures.

The gradual shift from 1/2-gram to 3-gram procedures supports the involvement of both transitional probability within adjacent syllables (Saffran et al., 1996) and the learning of lexical forms (Estes et al., 2007). The discussion of such dynamic procedural development will be elaborated in the general discussion section. The consistency of arriving at the 3-gram procedure further provides the basis for the robustness of behavioural results in statistical learning (with a moderate effect size for infants based on a meta-analysis, see Isbilen & Christiansen, 2022). Nevertheless, the limited trial proportions do not indicate consistent application of the learned procedure across trials. These results thus support the alternative argument of Alhama and Zuidema (2019) that continuous but incomplete learning of the pattern is sufficient for pattern differentiation.

Test phase. Given only two patterns in the test condition, in addition to the *n*-gram procedures, repetition can be detected objectively at the fourth position. Across conditions, the low-efficiency model only applies a low proportion of procedures that detect repetition at the third position (16.7%–17.8%, orange diamond), suggesting syllable omission prior to repetition detection. For the high-efficiency model, the 3-gram procedure (pale triangle) is applied more often in the word condition (33.2%) than in the part-word (10.7%) and non-word (11.4%) conditions, demonstrating the transfer of the learned 3-gram procedure to the consistent condition. The model favours the 1-gram procedure (54.3%, pink hourglass) in the non-word condition, and both the 1-gram (32.2%, pink hourglass) and 2-gram (35.4%, brown diamond) procedures in the part-word condition. This suggests that when the trained transitional probability of the pattern shifts from part-word to completely scrambled non-word, the application of single-syllable 1-gram procedures increases. The high-efficiency model occasionally correctly identifies syllable repetition at the fourth position (16.3%–23.1% across conditions, green square).

Taken together, the results show that the models can flexibly apply various procedures based on task changes, but the trial proportions of these procedures remain limited. The results further support both the transfer and readaptation of procedures (see Taatgen, 2013), which aligns with open procedural learning rather than all-or-none strategy acquisition (see Alhama & Zuidema, 2019).

4.2 Study 2: Syntactic algebraic processing

The simulated task is adapted from Marcus et al. (1999). The training includes the presentation of a series of trisyllabic patterns. However, these patterns are composed of variable tokens that follow a certain syntactic rule. For example, in *a-b-a*, class *a* and *b* may change constantly, but the identity repetition of the first and the third *a* remains the same.

Task material. The training phase presents *a-b-a* or *a-b-b* patterns. Specifically, class *a* is instantiated as "le," "wi," "ji," or "de," and class *b* is instantiated as "di," "je," "li," or "we," creating 16 possible trisyllabic patterns for each rule. The test phase presents two conditions, *c-d-c* and *c-d-d*, which are either consistent or inconsistent with the training patterns.[4] The syllable tokens in the test trisyllabic patterns do not overlap with those in

[4]Note that the original experiment used a within-subject design, while we adopted a between-subject design consistent with our approach for the lexical statistical learning task.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

the training patterns. Therefore, at the lexical level, the test patterns and the training patterns are completely different lexical forms. Specifically, class *c* is instantiated as "ba" or "ko," and class *d* is instantiated as "po" or "ga," creating 4 possible trisyllabic patterns for each rule. Note that the results corresponding to the two training conditions were pooled together in the original study, and we have followed the same approach in this study.

Presentation and processing rate. Each trisyllabic pattern is presented for 1500 ms, resulting in a syllable presentation rate of 500 ms per syllable. Additionally, there is a 1000-ms inter-pattern interval between the trisyllabic patterns. The model incorporates two efficiency levels, as described in Study 1, to simulate developmental trends. Each model run incorporated 100 training patterns and 10 test patterns. The simulation study included 100 separate model runs.

Simulated preferential focusing. Following Study 1, we analysed two dependent variables, namely overall on-task processing time and operator-level latency. We used linear regression to assess whether the two independent variables test condition (consistent or inconsistent) and model efficiency (high or low) influenced the on-task processing time and operator-level latency, again applying a forward-fitting model comparison procedure (see Simulated preferential focusing).

On-task processing time. While no significant improvement in model fit is observed when considering the main effects of test condition (compare Models 2 and 3; $F_{(1,797)} = 0.97$, $p = 0.32$, $\Delta AIC = 1.02$), a substantive enhancement in model fit occurs upon the inclusion of the model efficiency factor (compare Models 1 and 3; $F_{(1,797)} = 3084.60$, $p < 0.001$, $\Delta AIC = -1264.53$). The interaction is only improving the model fit marginally (compare Models 3 and 4; $F_{(2,796)} = 2.80$, $p = 0.09$, $\Delta AIC = -0.81$), thus we select Model 3 as the best-fitting model.

Model 3 shows no significant difference between consistent and inconsistent task conditions in either high- or low-efficiency simulations. Nevertheless, low-efficiency simulation exhibits longer on-task processing times ($\beta_{\text{low-high}} = 5.155$, $SE = 0.09$, $t = 55.54$, $p < 0.001$), indicating they are 5155 ms slower than high-efficiency simulations.

Context-based operator latency. The model fit is significantly improved by incorporating the main effects of test condition (compare Models 2 and 3; $F_{(2,797)} = 10.71$, $p < 0.001$, $\Delta AIC = -8.68$) and model efficiency (compare Models 1 and 3; $F_{(1,797)} = 1062.40$, $p < 0.001$, $\Delta AIC = -675.74$). After adding the interaction effect, model fit is, however, not improved (compare Models 3 and 4; $F_{(1,796)} = 2.35$, $p = 0.13$, $\Delta AIC = 0.36$). Thus, we choose Model 3 as the best-fitting model.

Model 3 reveals a significant difference in context-based operator latencies between consistent and inconsistent conditions across both high- and low-efficiency simulations ($\beta_{\text{con.-inc.}} = -0.0031$, $SE = 0.0009$, $t = -3.27$, $p = 0.001$). Specifically, the context-based operator latency is 3.1 ms faster in the consistent condition compared to the inconsistent condition. Operator latencies are 30 ms slower in low-efficiency simulations than in high-efficiency simulations ($\beta_{\text{low-high}} = 0.030$, $SE = 0.0009$, $t = 32.60$, $p < 0.001$).

Summary. In this study, on-task processing time and operator latency correspond to preferential focusing and immediate engagement, respectively. We observed differences only in operator latencies, not on-task processing times, between consistent and inconsistent conditions (see Figure 6). Thus, the simulated tendency of engagement corresponds to empirical literature indicating preferential focusing in algebraic tasks

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

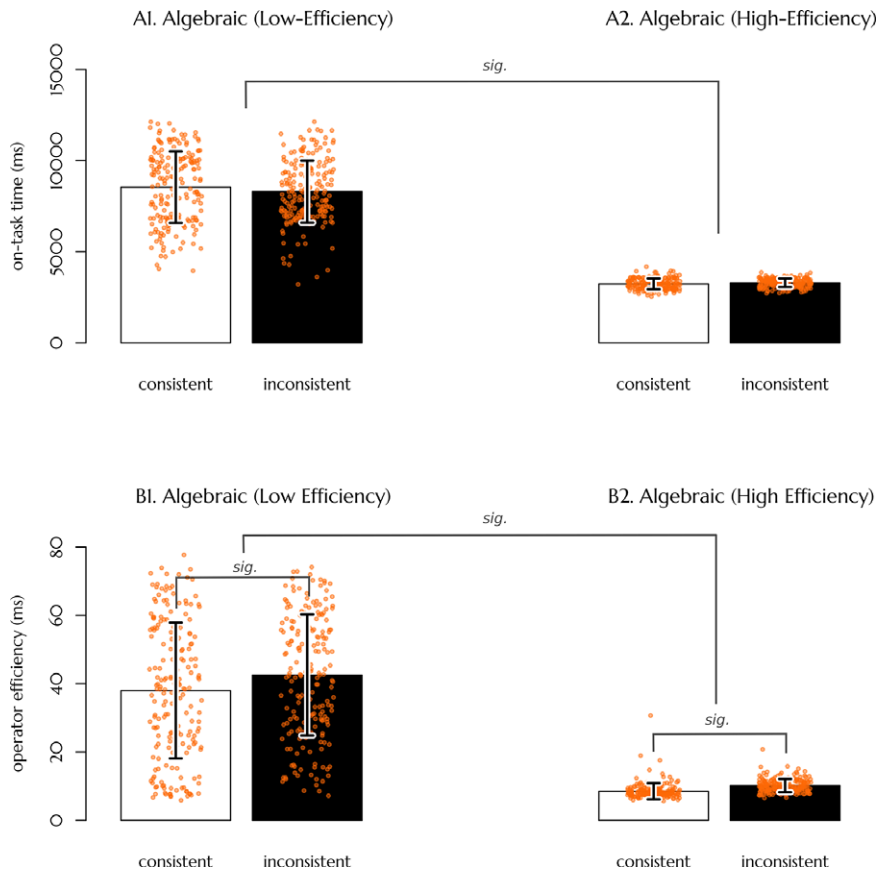


Figure 6. Simulated preferential focusing dynamics regarding algebraic tasks. (a) Averaged processing time under test conditions. Calculated as the sum of on-task operator latencies during the test phase. *Note*: y-axis: average processing time (in ms); x-axis: test conditions; white bars: consistent conditions; various gray bars: inconsistent conditions; error bars: ± 1 SD; orange dots: data points from individual model runs. (b) Averaged operator efficiency under test conditions. Calculated based on the average context-based latency (excluding fixed default action time) of all operators across model runs. *Note*: y-axis: average latency (in ms); x-axis: test conditions; white bars: consistent conditions; various gray bars: inconsistent conditions; error bars: standard deviations; orange dots: data points from individual model runs. The significance of regression coefficients is denoted by brackets and indicators (sig., $p < 0.001$). Note that the overarching brackets denote the main effect of model efficiency.

(Marcus et al., 1999; Rabagliati et al., 2018). In contrast, simulated preferential focusing is consistent with more recent replication studies that cast doubt on any effects of algebraic learning (Geambaşu et al., 2022; Visser et al., 2021). Similar to Study 1, higher-efficiency simulations lead to shorter on-task processing and operator latency. This is consistent with the general main effect on preferential focusing time (Dawson & Gerken, 2009; Frank et al., 2020).

Underlying procedural learning. For both Figures 7 and 8, figures (a) and (b) illustrate the performance of the low- and high-efficiency models. Appending number 1 depicts the proportion of procedures applied across learning blocks during the learning phase, while

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

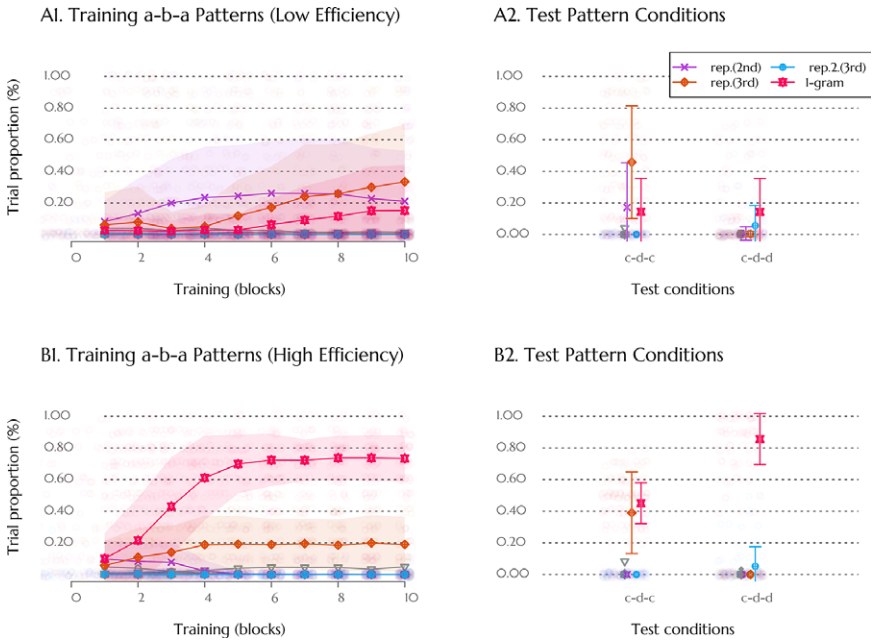


Figure 7. The averaged proportion of procedures applied by the model after training *a-b-a* across model runs. The first two repetition procedures detect repetition of input syllables compared to a already encoded working memory slot (slot 1) at the 2nd (purple cross) and 3rd (orange diamond) syllable positions, respectively. Repetition at the 2nd position is due to the omission of the middle token *d* in *c-d-c*. Another repetition procedure detects a match between the input and a different encoded working memory slot (slot 2) also at the 3rd syllable position (blue dot). Alternatively, the 1-gram procedure detects differences between the working memory pattern and the retrieved pattern immediately at the 2nd position (pink hourglass). *Note.* y-axis: averaged trial proportion of each procedure (within each block); training-phase x-axis (A1/B1): 10 blocks each consisting of 10 trisyllabic patterns; test-phase x-axis (A2/B2): test conditions; error band/bars: ±1 SD; transparent dots: data points from individual model runs. The sum of the trial proportions is not equal to 1, as the model may not use any procedure or may use more than one procedure in a trial.

number 2 depicts the proportion of procedures applied in a single block under various test conditions following the training phase. Note that figure (a1)/(b1) depicts the training phase preceding the *c-d-c* test condition only.

Training phase. *a-b-a training* (see Figure 7). Both low- and high-efficiency models learn 1-gram and repetition procedures suitable for processing *a-b-a*. The low-efficiency model reaches a higher proportion for the repetition procedure (33.4%, orange diamond) than the 1-gram procedure (15.2%, pink hourglass; see (a1)). In contrast, the high-efficiency model reaches a higher proportion for the 1-gram procedure (73.4%, pink hourglass) than the repetition procedure (19.0%, orange diamond; see (b1)). Only the low-efficiency model detects repetition at the second position due to syllable omission (21.1%, purple cross; see (a1)). *a-b-b training* (see Figure 8). Both low-efficiency and high-efficiency models learn only very limited procedures, with the exception that the high-efficiency model learns a high proportion of the 1-gram procedure. The low-efficiency model reaches a slightly higher proportion for the repetition procedure (10.7%, blue dot) than the 1-gram procedure (9.7%, pink hourglass; see (a1)). In contrast, the high-efficiency model reaches a high proportion for the 1-gram procedure (98.2%, pink hourglass) and a low proportion for the repetition procedure (1.1%, blue dot; see (b1)).

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

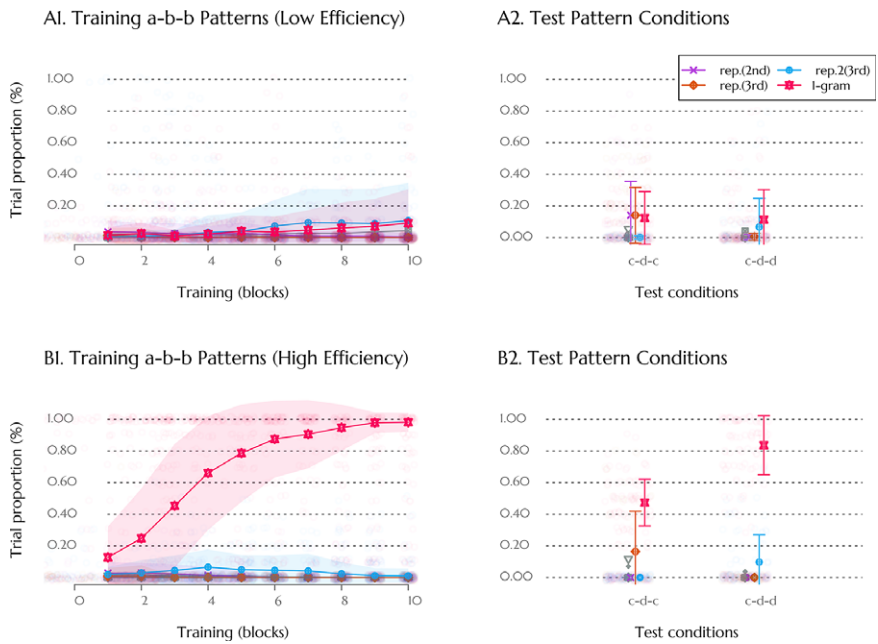


Figure 8. The averaged proportion of procedures applied by the model after training *a-b-b* across model runs. The first two repetition procedures detect repetition of input syllables compared to an already encoded working memory slot (slot 1) at the second (purple cross) and third (orange diamond) syllable positions, respectively. Repetition at the second position is due to the omission of the middle token *d* in *c-d-c*. Another repetition procedure detects a match between the input and a different encoded working memory slot (slot 2) also at the third syllable position (blue dot). Alternatively, the 1-gram procedure detects differences between the working memory pattern and the retrieved pattern immediately at the second position (pink hourglass). *Note.* y-axis: averaged trial proportion of each procedure (within each block); training-phase x-axis (A1/B1): 10 blocks each consisting of 10 trisyllabic patterns; test-phase x-axis (A2/B2): test conditions; error band/bars: ±1 SD; transparent dots: data points from individual model runs. The sum of the trial proportions is not equal to 1, as the model may not use any procedure or may use more than one procedure in a trial.

Based on the resource availability perspective (Taatgen et al., 2021), the low-efficiency model processes task patterns partially, allowing only the encoding of the current syllable and preventing immediate comparison, or completely ignoring the syllable. The less efficient model thus suppresses the lexical 1-gram procedure, which requires an additional comparison operator after encoding during a single-syllable presentation window. However, it can still encourage repetition procedures requiring only one input comparison operator without input encoding. Conversely, a high-efficiency model can quickly apply the slightly more complex lexical 1-gram procedure thanks to sufficient temporal resources.

Despite simulating preferential focusing dynamics at the operator efficiency level, the model results at both efficiency levels do not support the rule-based interpretation of Marcus et al. (1999). The low-efficiency model exhibits plural and limited procedural learning, whereas the high-efficiency model instead learns the alternative 1-gram procedure that fails to infer the syntactic regularity. Neither model demonstrated consistent repetition procedures. Our simulation results thus confirm the less robust findings of algebraic performance (small effect, see Rabagliati et al., 2018). Nevertheless, this still

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

supports the alternative hypothesis that continuous, albeit incomplete, learning is sufficient for producing the reported focusing dynamics (Alhama & Zuidema, 2019).

Test phase. *a-b-a training* (see Figure 7). The low-efficiency model performance shows that the trained repetition procedure (orange diamond) is applied more frequently in the consistent *c-d-c* condition (45.8%) but is not observed in the inconsistent *c-d-d* condition (0.0%). While the 1-gram procedure is suppressed in both *c-d-c* and *c-d-d* conditions (14.5/14.4%, pink hourglass), the inconsistent *c-d-d* condition also does not reveal readaptation of another repetition procedure (5.7%, blue dot). The high-efficiency model performance shows that the trained repetition procedure (orange diamond) is enhanced in the consistent *c-d-c* condition (24.4%) but is not found in the inconsistent *c-d-d* condition (0.0%). The repetition procedure suppresses the trained 1-gram procedure in the consistent *c-d-c* condition (45.1%) compared to the inconsistent *c-d-d* condition (85.7%). The model, however, does not exhibit readaptation of the alternative repetition procedure (5.2%, blue dot) in the inconsistent *c-d-c* condition. *a-b-b training* (see Figure 8). The low-efficiency model performance shows that both repetition (6.7% in *c-d-d*, blue dot; 14.1% in *c-d-c*, orange diamond) and 1-gram (11.3/12.4% in *c-d-d*/*c-d-c*, pink hourglass) procedures are applied with a low proportion across conditions. In the *c-d-c* condition, the model occasionally detects repetition at the second position due to syllable omission (14.1%, purple cross; see (a1)). The high-efficiency model maintains a high proportion for the 1-gram procedure (83.7%, pink hourglass) and a low proportion for the repetition procedure (9.8%, blue dot), in the consistent *c-d-d* condition. In the inconsistent *c-d-c* condition, the 1-gram procedure is reduced (47.4%), with limited readaptation of the alternative repetition procedure (16.4%, orange diamond).

In the test conditions, the tested patterns are changed in syllable content but maintain/change abstract syntactic repetition patterns. This means the learned lexical 1-gram procedure would become unsuitable. The shift in test patterns suppresses the 1-gram procedure in the low-efficiency model. After *a-b-a* training, this encourages the transfer of the repetition procedure to the consistent test condition, despite limited readaptation of the alternative repetition procedure in the inconsistent condition. However, the limited procedural learning of *a-b-b* training leads to very limited application of the repetition and 1-gram procedures in both test conditions. Note that after both training scenarios, the low-efficiency model may misinterpret the *c-d-c* condition as *c-c* by omitting the middle syllable. This then prompts the detection of second-position repetition. Likewise, the shift in test patterns also suppresses the 1-gram procedure in favour of the repetition procedure in the high-efficiency model, although only in consistent conditions. Regarding the procedures learned during the test phase, the simulation results present a picture of how algebraic patterns may be processed pluralistically. The results altogether support the PRIMs theory (Taatgen, 2013) concerning the flexible transfer and readaptation of procedures, and the argument favouring continuous procedural learning instead of an all-or-none strategy acquisition (see Alhama and Zuidema, 2019).

4.3 Study 3: Word-level phonological learning

This study applies the model to more naturalistic contexts, providing a proof-of-concept for developmental simulation. The material is based on the phoneme sequences extracted from the CHILDES infant-directed corpus (MacWhinney, 2000). Note that although each word in a sentence utterance has a varying number of phonemes or *phoneme length* (e.g., the word-level utterance "Charlie's" has a length of six phonemes, namely "CH," "AA1," "R," "L," "IY0," and "Z"), these words are embedded into a continuous syllable stream

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

without clear word boundaries. The objective is to investigate whether the model learns word-level phoneme/phonological patterns (e.g., "CH_AA1_R_L_IY0_Z") at different rates when adjusting the assumed developmentally related action time.

Task material and presentation rate. The training material is obtained from 14 hours of recordings of a mother's speech towards an infant from 6 to 10 months (Soderstrom et al., 2008). Mother's sentence-level utterances were transcribed into individual phonemes using the CMU dictionary. Utterance boundaries are represented by four consecutive phoneme-level "#" symbols. These phonemes were then unlisted to create moving windows, each consisting of three phonemes. The moving window sequence was uninterrupted, each lasting for 50 ms (based on the typical phoneme presentation rate, see Menn et al., 2023). This continuous sequence is divided into 50 blocks (i.e., each contains 2263 windows).

Model adjustment. The primitive operators used in this study are a subset of those applied in Studies 1 and 2. This simplification follows from Study 1 results, where *n*-gram procedures are capable of learning multisyllabic patterns. We focus on such procedures exclusively, without considering alternative repetition procedures and operators that are irrelevant for lexical learning. Specifically, we exclude repetition-based match-detection operators, along with other comparison operators (e.g., mismatch between input and working memory, or match between working memory and declarative memory) that do not lead to task recognition. For encoding triphoneme moving windows, the model now simultaneously encodes a moving window (e.g., "CH_AA1_R") into three working memory slots (e.g., slots 1, 2, and 3). To accommodate longer word-level phoneme sequences, we have increased the number of working memory slots (now up to the ninth slot) an operator can encode maximally.

Processing rate. To model developmental differences in word learning, we adjust the fixed action times, including a faster range (60, 80, and 100 ms) and a slower range (200 and 300 ms). We hypothesised that the current model can still learn the phoneme patterns when omitting three triphoneme windows (e.g., the first window "CH_AA1_R" and the fourth window "L_IY0_Z" into slots 1, 2, and 3 and slots 4, 5, and 6 to learn "Charlie's") without needing to successively encode all consecutive moving windows. In other words, the current model can learn the pattern when the minimum operator latency is below 200 ms (omission of three windows and a partial processing of the fourth window; with a slightly longer latency, the model would instead learn "Charie's"). Note that our simulation specifically focuses on the assumed developmentally-relevant processing rate. For convenience, all models underwent equivalent training on approximately 14 hours of material, regardless of any prior learning history an actual child may have had.

Learning outcomes and trajectories. Linear regression analyses are conducted to investigate the relationship between the activation of word-level phonological patterns[5]

[5]This study utilizes ACT-R's optimized learning equation for base-level activation (Anderson, 2007) to inform the activation of phonological patterns (i.e., activation of chunks/declarative items). The equation $\log\left(\text{fixed activation} + N \cdot (\Delta t)^{-d}/(1-d)\right)$ simplifies to $\log(N)$ when fixed activation and decay parameter d are set to 0. Note that the ACT-R decay parameter is most effective for short-interval forgetting (0–10 min) but becomes overly pessimistic for long-interval forgetting (van der Velde et al., 2022). Additionally, we do not scale the fixed activation, leaving the activation level arbitrarily valued. This simplification makes declarative item activation solely determined by the frequency factor N. In standard PRIMs, N is referenced by the number of pattern segmentations, where the specific pattern is placed from the working memory

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

(dependent variable) and two independent variables: word frequency and phoneme length. In this study, we only consider the learned phonological patterns that correspond to the actual words (or word-level phonological patterns) applied in the training material. Word frequency is defined as the actual frequency of each word within the entire trained task material. Phoneme length, as described previously, represents the number of phonemes embedded in a word-level phonological pattern (e.g., the word "Charlie's" has a length of six phonemes). Separate linear regression analyses are conducted for each model efficiency level (i.e., fixed action time). Within each level, we run two distinct models: one to assess the effect of word frequency on word activation and another for phoneme length. Simulation results first revealed that enhanced model efficiency (characterised by a decreased action time range, from 200–300 ms to 60–100 ms) led to the learning of more word-level phonological patterns (63 and 91 patterns, or 99, 76, and 261 patterns). From Figure 9a, we can also see that when model efficiency no longer supports learning phonological patterns beyond a single triphoneme window (i.e., slower than 200 ms), the model no longer learns phonological patterns beyond three phonemes (fixed action time of 300 ms, blue dot). Conversely, when increasing model efficiency, the model learns longer phonological patterns (reaching five phonemes, fixed action time of 60 ms, orange dot). Regression analyses indicated that higher word frequency is associated with increased acquired pattern activation (300 ms, β=0.51; 200 ms, β=0.43; 100 ms, β=0.57; 80 ms, β =0.49; 60 ms, β =0.54; ps < 0.01), while longer phoneme length is associated with decreased acquired pattern activation (300 ms, $\beta = -0.28$, $p = 0.03$; 200 ms, $\beta = -0.39$, p<0.001; 100 ms, $\beta = -0.19$, $p = 0.07$; 80 ms, $\beta = -0.26$, $p = 0.03$; 60 ms, $\beta = -0.23$, p<0.001). Note that the activation value is collected at the end of the 50th learning block and that in all regressions, activation, frequency, and phoneme length values are normalised.

Previous large cross-linguistic findings suggest a strong effect of frequency on both word comprehension and production (Braginsky et al., 2019). This link is evident in our model, given that the frequency of pattern segmentation is directly linked to pattern activation. Besides, a moderate effect of the number of phonemes on word production, instead of word comprehension, is also identified across different languages (Braginsky et al., 2019). Consequently, the number of phonemes may be associated with phonological learning, assuming production reflects phonological competence (see Fikkert, 2007). As discussed, the learning of longer patterns is related to model efficiency relative to the triphoneme moving windows. Thus, the simulated results, at a procedural level, may also provide insights into why the phonological sequence of longer words is more difficult to learn for younger or linguistically delayed children. Our model nevertheless cannot provide explanations for more semantic-related factors such as babiness and concreteness (Braginsky et al., 2019).

Linear regression analyses are also conducted to investigate activation change of word-level phonological patterns over the training blocks. In this case, the block numbers (1 to 50 blocks) serve as the independent variable. Separate linear regression analyses are again conducted for each model efficiency level (i.e., fixed action time). The model shows training-related enhancement of average pattern activation over the training blocks (slow range: 300 ms, $\beta = 0.007$; 200 ms, $\beta = 0.006$; moderate rage: 100 ms, $\beta = 0.013$; 80 ms, $\beta = 0.012$; fast range: 60 ms, $\beta = 0.031$, ps < 0.001). To determine if training-related enhancement progresses in discrete steps across model efficiency ranges, we combined data from all

module to the declarative module. In this article, we do not explore whether the learning-curve-like logarithm function could be alternatively explained by a contextual learning mechanism.

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

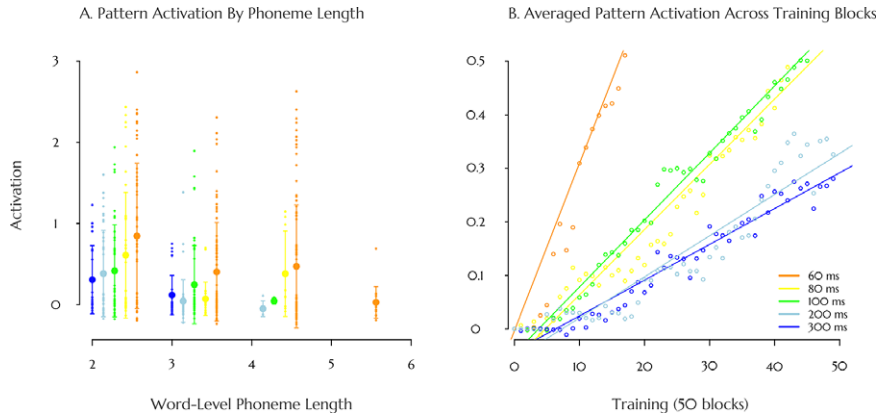


Figure 9. Learning outcomes and trajectories of word-level phonological patterns. (a) Acquired activation of word-level phonological patterns at different model efficiency levels at the end of the 50th block. *Note*: y-axis: activation of phonological patterns (i.e., chunk activation in ACT-R); x-axis: phoneme length of word-level phonological patterns. The color coding from blue to orange indicates increasing efficiency from 60 to 300 ms (i.e., default action time). Each mini-dot represents a word-level phonological pattern. The means and standard deviations are highlighted by the larger dots and error bars. (b) The trajectory of word-level phonological patterns over the 50 blocks at different model efficiency levels. *Note*: y-axis: activation of phonological patterns (i.e., chunk activation in ACT-R); x-axis: 50 blocks; color coding as above. The dots represent the averaged phonological activation across the blocks. They are either averaged from the shared patterns within the high (60–100 ms) or low (200–300 ms) efficiency range.

efficiency levels and conducted pairwise contrast analyses on the regression coefficients (emmeans package, Lenth, 2020). As Figure 9b illustrates, pairwise contrasts of regression coefficients reveal that training-related enhancement increases with model efficiency, moving from slow (200–300 ms) to moderate (80–100 ms) and then fast (60 ms) ranges (slow to moderate range, $betas_{\text{m.-s.}} = [0.0048, 0.0058]$, $SEs = 0.0006$, $ts = [7.69,9.29]$; moderate to fast, $betas_{\text{f.-m.}} = [0.0187, 0.0190]$, $SEs = 0.0006$, $ts = [29.76,30.28]$; $ps < 0.0001$). However, regression coefficients within the same model efficiency range do not show significant differences ($ps < 0.05$). The significant differences in training-related activation enhancement between model efficiency ranges demonstrate young children's capacity for word learning even when their neural response rates (corresponding to model efficiency levels) are considerably slower than the phoneme presentation rate (see Menn et al., 2023). Furthermore, the lack of significant differences in training-related activation enhancement within each model efficiency range may explain the limited age-related differences observed in statistical learning (see Isbilen & Christiansen, 2022) during developmental stages when brain maturation results in a clear improvement of processing efficiency (see Dubois et al., 2014). Note, however, that our model only considers triphoneme features for demonstration purposes. In reality, a child may also capture other slow-presenting, overarching features (e.g., the voicing, sonorant, and continuant properties of words; see Menn et al., 2023), other than the faster-presenting syllable/phonemes.

5. General discussion

In the following discussion, we explore the interconnected levels reflected in our simulations of lab-based phenomena and naturalistic word learning. These include the

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

behavioural, procedural, and language development levels. At the behavioural level, we consider whether the simulation results align with the observed preferential focusing dynamics in lab-based studies and whether such behavioural outcomes guarantee the explicit acquisition of strategies. At the procedural level, we examine how the various phenomena of statistical and algebraic learning can be interpreted through differential procedural learning within a unified cognitive architecture. We then briefly discuss the potential of such a framework to connect phenomena across different language levels (see Benders & Blom, 2023). Within the procedural level, we further discuss the dynamic development from partial to more refined procedures and how diverse procedures can emerge within a single task. At the more naturalistic level of age-related language development, the study employs phonological learning as a proof-of-concept to demonstrate how our model can contribute to understanding both typical and atypical language development.

5.1 Preferential focusing and implicit procedural learning

Alhama and Zuidema (2019) argue that preferential focusing dynamics reflect an ongoing learning process rather than all-or-none strategy acquisition. Our findings support this interpretation. While our model simulated preferential focusing dynamics, especially in high-efficiency conditions, both training and test performance in the two lab-based phenomena did not consistently reflect procedural application across trials. Instead, procedural application was either limited in terms of trial proportion (in statistical learning task performance) or led to unexpected procedures (in algebraic task performance) during training. Additionally, test performance did not definitively indicate strategy success or failure but rather demonstrated flexibility. Simulations revealed the model's ability to transfer learned procedures to consistent test scenarios and rapid re-adaption to inconsistent ones. Therefore, our simulation results do not support the hypothesis that preferential focusing dynamics imply strategy acquisition. In fact, preferential focusing dynamics may result from a different pattern of strategy development than expected. One example is the algebraic performance in high-efficiency models representing older infants (action time = 500 ms), which produced mostly lexical learning, contradicting the interpretation that differential looking patterns would signal algebraic or syntactic learning (Marcus et al., 1999). However, this procedural learning outcome resulted in a directional bias (based on average operator latency) that was consistent with the original findings. Nonetheless, this directional bias is not robust, as evidenced by the extremely subtle differences in both macro-level on-task focusing time and micro-level operator-level latency. These simulated findings may then explain why it is difficult to replicate the preferential dynamics in algebraic tasks, even with high statistical power (Geambaşu et al., 2022; Rabagliati et al., 2018; Visser et al., 2021). Taken together, preferential focusing dynamics may be better reflected as a continual learning process, where any learned pattern, whether partial or unexpected, leads to differentiation of the various test conditions.

Further regarding the directional bias of preferential focusing, the Hunter–Ames' model suggests that younger infants may have difficulty forming task representations, even with sufficient trials. This can lead to a preference for familiarity, as they are less likely to disengage from learning the tasks. In contrast, older children often exhibit a novelty preference, disengaging from learned tasks more quickly and focusing on new ones (see Hunter & Ames, 1988). Our simulated on-task processing time results, in numerical values, align with these age-related preferential directions. By modifying the

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

developmental aspect of operator latency, models representing the performance of younger infants (action time = 300 ms) display a familiarity preference in algebraic processing but no preference in statistical learning. In contrast, models representing the performance of older infants (action time = 50 ms) show a novelty preference in both statistical learning and algebraic tasks. Our overall simulation results are similar to the anticipated directional trends of Hunter and Ames (1988). Nevertheless, the explanation of preferential focusing dynamics differs between the Hunter–Ames' model and strategy/processing-based accounts (see Introduction). Moreover, recent studies have challenged the overall U-shaped trend proposed by Hunter and Ames (1988). This is evidenced by meta-analyses demonstrating less consistent directional effects (Black & Bergmann, 2017; Isbilen & Christiansen, 2022; Rabagliati et al., 2018).

One reason for this discrepancy is that infants may use different available strategies for a given task (Houston-Price & Nakai, 2004). For instance, algebraic tasks are complicated by multiple interpretations (Gerken, 2006, 2010). Furthermore, the meaningfulness of the task features also moderates preferential dynamics (Rabagliati et al., 2018). For the statistical learning task, while laboratory tasks using pseudo-words show an overall novelty preference (Black & Bergmann, 2017), more naturalistic word learning (real versus pseudo-words) consistently reveals a reversed familiarity preference (Bergmann & Cristia, 2015). In naturalistic settings, familiarity preference may arise because infants continue to extract information from real words or sentences. In contrast, a less meaningful task leads to more rapid disengagement and a preference for novelty, as there is no further information to be gained from the task (see footnote 3, Bergmann & Cristia, 2015). Note that familiarity preference is not limited to naturalistic word learning (see Bergmann & Cristia, 2015) but has also been observed in lab-based studies with task materials embedded in meaningful frames (e.g., Saffran, 2001) or in cross-situational learning paradigms where lexical inference is required (e.g., Smith & Yu, 2008). Clarifying "general information processing characteristics in infants" (Bergmann & Cristia, 2015) is thus essential for a better understanding of their focusing preferences. In this study, we show that the procedures openly discovered by the model are diverse and may contain operator sequences of different lengths (see Figures 5 and 8), which further influences processing duration and focusing time.

In contrast, when we would assume that the strategy/procedure that infants use is consistent, the U-shaped trend may be preserved. This is, for instance, revealed in a task where the familiar and novel test conditions are equivalent in task complexity (Kosie et al., 2023). In this respect, the motivational account of Hunter and Ames (1988) can be interpreted based on contextual learning within a resource availability framework (Taatgen et al., 2021, see Introduction). Briefly, initially insufficient learning corresponds to low information processing efficiency. At this point, the model gradually moves towards sufficient processing of the task. During this phase, the length of the operator sequence gradually increases, leading to an increasing focus on the current task (familiarity preference). However, when the processing efficiency of the appropriate procedure continues to increase, the overall on-task processing time decreases, leading to disengagement from the current task and longer engagement with a novel task (novelty preference). Note that the overall processing efficiency of the model, corresponding to a developmental factor, also moderates this trend. Importantly, this U-shaped trend (Hunter & Ames, 1988) hinges on the assumption that the procedures (or strategies) acquired during different test conditions are consistent, which is clearly not the case in our simulation results. This echoes the recent evidence that casts doubt on the prediction of Hunter and Ames' model (e.g., see Geambaşu, 2018; Geambaşu et al., 2022; Raz et al., 2023; Visser et al., 2021). The moderation of multiple factors on preferential focusing,

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

including the trial-by-trial trend of focusing on familiar and novel tasks, has been investigated in greater detail in a separate simulation study.

5.2 Open procedural learning within unified architecture

Our model demonstrates the potential to unify theoretical accounts across linguistic phenomena. We show how different patterns, such as statistical learning and algebraic tasks, can be gradually recognised from the bottom-up. This involves gradually inferring procedures, resembling the hypothesis space of rules (see Frank et al., 2016; Frank & Tenenbaum, 2011). Rather than predefining these rules, they are assembled from a constrained set of processing elements through trial and error (inspired by de la Cruz-Pavía & Gervain, 2021, in this study). The model can flexibly discover *n*-gram procedures, similar to memory-based models (see French et al., 2011; Mareschal & French, 2017; Thiessen & Pavlik, 2012). It can also apply repetition procedures related to processing that does not require long-term memory. Our simulation study revealed lexical *n*-gram procedures in the statistical learning task (e.g., Estes et al., 2007) and the plural interpretation in the algebraic task. (see Gerken, 2006, 2010). Moreover, the time constraints imposed by the model's efficiency levels contribute to the simulation of developmental trends in the two targeted phenomena. Only high-efficiency models enable lexical learning, facilitated by their temporal resource to make declarative retrieval (i.e., *n*-gram procedures). However, suppressing lexical interpretation benefits repetition detection, which relies solely on input working memory comparison. When considered from a developmental perspective, the acquired underlying procedures further demonstrate why statistical learning likely has a much later onset compared to algebraic processing (see Bergmann & Cristia, 2015; de la Cruz-Pavía & Gervain, 2021; Wilson et al., 2018).

Overall, our simulation findings support the efficacy of *n*-gram procedures that favour chunking- and memory-based mechanisms in linguistic acquisition (see Isbilen & Christiansen, 2022). This core memory-based mechanism (i.e., chunking and declarative retrieval) aligns with other cognitive architecture-based models that simulate real-life grammar learning phenomena, such as reflexive-pronoun inference (van Rij et al., 2010) or past-tense learning (Taatgen & Anderson, 2002). The open procedural learning within a unified computational framework described in this article allows for the reconstruction of these models (among other cognitive architectural models that previously relied on predefined production rules) when supplemented with the necessary primitive operators. Furthermore, while this article focused on simple phenomena at the level where operators form a certain procedure, the framework can be leveraged across language levels (as mentioned in Benders & Blom, 2023). For instance, we have just mentioned the additional requirement of the lexical inference process during naturalistic or cross-situational word learning (Bergmann & Cristia, 2015; Smith & Yu, 2008). To model these more complex language learning phenomena, the model needs to move beyond single-procedure learning and instead acquire procedure-procedure associations, utilising a similar contextual learning mechanism. The model's capacity to address other more complex language phenomena has been detailed elsewhere (Ji et al., 2025a, b).

5.3 Emergence of procedures over learning phase

Our study also indicates that young children may not consistently employ a single procedure to process patterns but may gradually transition between procedures as they

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

become more familiar with the task. For example, in statistical learning, there can be a progression from learning adjacent syllables (Saffran et al., 1996) to lexical forms (Estes et al., 2007), whereas in algebraic processing, the model may develop a procedure from two equally plausible interpretations (Gerken, 2006, 2010).

The dynamic development and the gradual emergence of procedures may provide a resolution to the debate regarding whether early language is based on chunking of small tokens to form word-level patterns (e.g., Saffran & Kirkham, 2018) or segmenting from larger utterance boundaries to form smaller units (e.g., Arnon, 2021). Initially, when encountering an unfamiliar pattern, whether statistical or algebraic, the model treats it concretely as a lexical pattern and encodes as many syllable tokens as possible. This process generates random *n*-grams of varying lengths, which are gradually stored in long-term memory for future reference. This is consistent with the findings that young children initially segment speech based on larger utterance boundaries (Soderstrom, 2003), rather than their smaller units. This very initial phase, therefore, aligns with the "start big" theory (Arnon, 2021), where the presented utterance is gradually segmented and placed into memory.

However, when the learned longer patterns are incomplete or inaccurate due to processing limitations or syllable omissions, the model may be restricted to predicting any following syllable (1-gram procedure) or only the next syllable (2-gram procedure). This marks the stages of learning transitional probabilities between adjacent syllables. As training progresses, the model then accurately identifies the lexical boundary of the trisyllabic pattern (3-gram procedure). When the transitional period is prolonged, the pattern inference phase aligns more closely with the chunking perspective (Saffran & Kirkham, 2018). Nevertheless, when the model is sufficiently efficient at rapidly identifying lexical forms, the transitional period becomes so brief that it is difficult to detect (see Figure 5). This learning trajectory would once again support the "start big" perspective (Arnon, 2021). Supporting this view, a meta-analysis has revealed that the strength of transitional probability is not as decisive in determining the extent of auditory statistical learning as previously thought (Isbilen & Christiansen, 2022).

The dynamic perspective is also applicable to the algebraic task. In the initial phase, the model possesses no declarative knowledge and would therefore be incapable of pattern inference. Consequently, the model initially only identifies repetitions between the input and the encoded working memory content (repetition procedure). In a later phase, where declarative retrieval is possible, the model would then start to infer the lexical pattern. Due to the variability of the token classes, however, the model cannot predict any adjacent syllable (1-gram procedure). This mismatch detection would suppress repetition detection as anticipated by Marcus et al. (1999). Therefore, the model would detect repetition only during the initial learning phase but would rapidly transition to *n*-gram procedures if declarative retrieval becomes possible. Our study thus highlights the dynamic nature of young children's learning. Their learning mechanism may evolve over time, shifting from initial segmentation (see Christiansen et al., 1998) or repetition (see repetition rule in Frank & Tenenbaum, 2011) to later memory-based chunking (see French et al., 2011; Perruchet & Vinter, 1998) or memory-based pattern mismatch (see French et al., 2011; Thiessen & Pavlik, 2012). Therefore, young children's learning process cannot be restricted to a single computational model.

Furthermore, our simulations of lab-based phenomena may provide implications for the changing cognitive processes involved in more naturalistic language learning (e.g., cross-situational word learning). Although our research only examined phonological learning, semantic aspects are readily anticipated. When the model has not formed stable

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

lexical/phonological patterns, semantic inference would be unstable. Similar to the findings in Estes et al. (2007), infants struggle to learn the relationship between novel non-word/part-word phonological patterns and corresponding image referents. The model therefore implies that the sufficient learning of phonological patterns (referred to as procedural learning in Walker et al., 2020) precedes the semantic pairing of the referent (referred to as declarative learning in Walker et al., 2020), aligning with recent results in cross-situational word learning. Specifically, using correlational analyses between cross-situational word learning performance and cognitive measures, Walker et al. (2020) found that initial word learning performance (e.g., learning nouns, adjectives, and markers) is related to phonological learning (serial reaction performance), whereas later performance (second-day follow-up) is related to semantic inference (pair association performance). Thus, the anticipated performance of our model and existing literature indicate that phonological/procedural learning precedes semantic/declarative learning (Goffman & Gerken, 2023), rather than the reverse (compare Ullman et al., 2020).

5.4 Procedural learning and language development

To demonstrate the current model's relevance to naturalistic language learning, this simulation study focuses on word-level phonological learning from infant-directed speech as a proof of concept. Learning the phonological form of words is a crucial aspect of language acquisition. In Study 3, we demonstrate how a model with a slower processing rate can still learn phonological patterns, addressing a question raised by Menn et al. (2023). By progressively increasing the model's efficiency levels, the model correctly segments longer word-level phonological patterns at the end of training and exhibits faster enhancement of corresponding phonological activation over the training blocks. Moreover, analyses of simulated results further support the relevance of word frequency and phoneme length in word learning outcomes (see Braginsky et al., 2019).

In addition to typical development, phonological learning is also implicated in developmental language disorder (DLD, Bishop et al., 2017). Children with DLD exhibit developmental delays, primarily characterised by difficulties with the encoding of syllable and prosody patterns, while other aspects of word learning, such as semantic association, remain unaffected (Goffman & Gerken, 2023; Ullman et al., 2020). However, recent studies suggest that while DLD children may not be impaired in all procedural abilities, they struggle specifically with those requiring the organisation of sequences, both linguistic-specific and beyond (Goffman & Gerken, 2020, 2023). Although our analysis did not include cases of phoneme omission or substitution (such as producing "Charie" instead of "Charlie"), the model's developmentally induced encoding errors are implicated in the reduced ability to correctly segment longer word-level phonological patterns. These results demonstrate the crucial role of encoding in phonological learning, offering a procedural interpretation of DLD.

Furthermore, our model aligns with the notion that procedural learning typically begins before declarative knowledge acquisition. This finding is consistent with treatment studies involving individuals with DLD. For example, learning materials that implicitly emphasise particular grammatical patterns of word-meaning relationships can help children with DLD, whereas direct instruction may impede their learning (e.g., Ferman et al., 2019; Plante & Gómez, 2018). While semantic processes are not included in this study, they would likely require extra temporal resources to process. For example, van Rij

https://doi.org/10.1017/S0305000925100159 Published online by Cambridge University Press

et al. (2010) demonstrated that a slower presentation pace was effective in helping children acquire reflexive pronouns, especially those struggling with faster rates. Nevertheless, explanations for processing efficiency, both developmental (i.e., myelination) and experience-based (i.e., learning), rely on the distributed connections between cortical areas (represented by cognitive modules in our model). This distributed nature may explain why DLD cannot be pinned down to a single brain structure or function (see Bishop et al., 2017). Consistent with the contextual learning account applied in this study, recent studies have identified the corticostriatal circuit as relevant for DLD (Abbott & Love, 2023; Ullman et al., 2024).

6. Conclusion

This simulation study proposes a learning-driven cognitive architecture that models representative lab-based statistical learning and algebraic processing phenomena, as well as naturalistic phonological learning from child-directed speech. Our analysis examines behavioural preferential focusing, underlying procedural learning, and the dynamic evolution of procedures in lab-based tasks, along with age-related trends in both lab-based and naturalistic word learning. These findings support an implicit procedural learning perspective, suggesting that the specific task (lexical or syntactic) is less critical than the task-induced distributional model contexts that dynamically shape information processing and procedural representation.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0305000925100159.

Acknowledgements. The authors would also like to thank action editor Dr. Titia Benders and two anonymous reviewers for their invaluable feedback and comments on earlier versions of this article.

Author contribution. Conceptualisation: YJ, JvR, NT; Methodology: YJ; Software – Model Scripts: YJ; Software – Source Development: NT, YJ; Formal analysis: YJ, JvR; Investigation: YJ; Writing – Original Draft: YJ; Writing – Review & Editing: YJ, JvR, NT; Visualisation: YJ; Supervision: JvR, NT.

Competing interests. The authors declare no conflict of interest.

Disclosure of use of AI tools. During the preparation of this work, the author(s) used Gemini to refine the language of this manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Abbott, N., & **Love, T.** (2023). Bridging the divide: Brain and behavior in developmental language disorder. *Brain Sciences*, **13**, 1606. https://doi.org/10.3390/brainsci13111606

Alhama, R. G., & **Zuidema, W.** (2018). Pre-wiring and pre-training: What does a neural network need to learn truly general identity rules? *Journal of Artificial Intelligence Research*, **61**, 927–946. https://doi.org/10.1613/jair.1.11197

Alhama, R. G., & **Zuidema, W.** (2019). A review of computational models of basic rule learning: The neural-symbolic debate and beyond. *Psychonomic Bulletin and Review*, **26**, 1174–1194. https://doi.org/10.3758/s13423-019-01602-z

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195324259.001.0001

Arnon, I. (2021). The starting big approach to language learning. *Journal of Child Language*, **48**, 937–958. https://doi.org/10.1017/s0305000921000386

Bates, E. (1979). The emergence of symbols: Cognition and communication in infancy. https://doi.org/10.1016/c2013-0-10341-8

Benders, T., & **Blom, E.** (2023). Computational modelling of language acquisition: An introduction. *Journal of Child Language*, **50**, 1287–1293. https://doi.org/10.1017/s0305000923000429

Bergmann, C., & **Cristia, A.** (2015). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, **19**, 901–917. https://doi.org/10.1111/desc.12341

Bishop, D. V., **Snowling, M. J.**, **Thompson, P. A.**, & **Greenhalgh, T.** (2017). Phase 2 of catalise: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, **58**, 1068–1080. https://doi.org/10.1111/jcpp.12721

Black, A., & **Bergmann, C.** (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In *39th annual meeting of the cognitive science society* (pp. 124–129). Cognitive Science Society.

Bouchon, C., **Nazzi, T.**, & **Gervain, J.** (2015). Hemispheric asymmetries in repetition enhancement and suppression effects in the newborn brain. *PLoS One*, **10**, e0140160. https://doi.org/10.1371/journal.pone.0140160

Braginsky, M., **Yurovsky, D.**, **Marchman, V. A.**, & **Frank, M. C.** (2019). Consistency and variability in children's word learning across languages. *Open Mind*, **3**, 52–67. https://doi.org/10.1162/opmi_a_00026

Chen, A., **Peter, V.**, & **Burnham, D.** (2016). Auditory ERP response to successive stimuli in infancy. *PeerJ*, **4**, e1580. https://doi.org/10.7717/peerj.1580

Christiansen, M. H., **Allen, J.**, & **Seidenberg, M. S.** (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, **13**, 221–268. https://doi.org/10.1080/016909698386528

Dawson, C., & **Gerken, L.** (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, **111**, 378–382. https://doi.org/10.1016/j.cognition.2009.02.010

de la Cruz-Pavía, I., & **Gervain, J.** (2021). Infants' perception of repetition-based regularities in speech: A look from the perspective of the same/different distinction. *Current Opinion in Behavioral Sciences*, **37**, 125–132. https://doi.org/10.1016/j.cobeha.2020.11.014

Dehaene, S., **Kerszberg, M.**, & **Changeux, J. P.** (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, **95**, 14529–14534. https://doi.org/10.1073/pnas.95.24.14529

Di Liberto, G. M., **Attaheri, A.**, **Cantisani, G.**, **Reilly, R. B.**, **Ni Choisdealbha, A.**, **Rocha, S.**, **Brusini, P.**, & **Goswami, U.** (2023). Emergence of the cortical encoding of phonetic features in the first year of life. *Nature Communications*, **14**, 7789. https://doi.org/10.1038/s41467-023-43490-x

Dubois, J., **Dehaene-Lambertz, G.**, **Kulikova, S.**, **Poupon, C.**, **Hüppi, P.**, & **Hertz-Pannier, L.** (2014). The early development of brain white matter: A review of imaging studies in fetuses, newborns and infants. *Neuroscience*, **276**, 48–71. https://doi.org/10.1016/j.neuroscience.2013.12.044

Emberson, L. L., **Misyak, J. B.**, **Schwade, J. A.**, **Christiansen, M. H.**, & **Goldstein, M. H.** (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science*, **22**, e12847. https://doi.org/10.1111/desc.12847

Endress, A. D., **Nespor, M.**, & **Mehler, J.** (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, **13**, 348–353. https://doi.org/10.1016/j.tics.2009.05.005

Estes, K. G., **Evans, J. L.**, **Alibali, M. W.**, & **Saffran, J. R.** (2007). Can infants map meaning to newly segmented words?: Statistical segmentation and word learning. *Psychological Science*, **18**, 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Ferman, S., **Kishon-Rabin, L.**, **Ganot-Budaga, H.**, & **Karni, A.** (2019). Deficits in explicit language problem solving rather than in implicit learning in specific language impairment: Evidence from learning an artificial morphological rule. *Journal of Speech, Language, and Hearing Research*, **62**, 3790–3807. https://doi.org/10.1044/2019_jslhr-l-17-0140

Fikkert, P. (2007). *Acquiring phonology* (pp. 537–554). Cambridge University Press. https://doi.org/10.1017/cbo9780511486371.024

Frank, M. C., **Alcock, K. J.**, **Arias-Trejo, N.**, **Aschersleben, G.**, **Baldwin, D.**, & **Sea, B.** (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, **3**, 24–52. https://doi.org/10.1177/2515245919900809

Frank, M. C., **Lewis, M.**, & **MacDonald, K.** (2016). A performance model for early word learning. *Proceedings of the 38th annual conference of the cognitive science society*, **38**, 2609–2614.

Frank, M. C., & **Tenenbaum, J. B.** (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, **120**, 360–371. https://doi.org/10.1016/j.cognition.2010.10.005

French, R. M., **Addyman, C.**, & **Mareschal, D.** (2011). Tracx: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, **118**, 614–636. https://doi.org/10.1037/a0025255

Frost, R., **Armstrong, B. C.**, & **Christiansen, M. H.** (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, **145**, 1128–1153. https://doi.org/10.1037/bul0000210

Geambașu, A. (2018). *Simple rule learning is not simple. Studies on infant and adult pattern perception and production.* PhD thesis, Leiden University.

Geambașu, A., **Spit, S.**, **van Renswoude, D.**, **Blom, E.**, **Fikkert, P. J.**, **Hunnius, S.**, **Junge, C. C.**, **Verhagen, J.**, **Visser, I.**, **Wijnen, F.**, & **Levelt, C. C.** (2022). Robustness of the rule-learning effect in 7-month-old infants: A close, multicenter replication of Marcus et al. (1999). *Developmental Science*, **26**, e13244. https://doi.org/10.1111/desc.13244

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, **98**, B67–B74. https://doi.org/10.1016/j.cognition.2005.03.003

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, **115**, 362–366. https://doi.org/10.1016/j.cognition.2010.01.006

Gervain, J., **Macagno, F.**, **Cogoi, S.**, **Peña, M.**, & **Mehler, J.** (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, **105**, 14222–14227. https://doi.org/10.1073/pnas.0806530105

Goffman, L., & **Gerken, L.** (2020). An alternative to the procedural-declarative memory account of developmental language disorder. *Journal of Communication Disorders*, **83**, 105946. https://doi.org/10.1016/j.jcomdis.2019.105946

Goffman, L., & **Gerken, L.** (2023). A developmental account of the role of sequential dependencies in typical and atypical language learners. *Cognitive Neuropsychology*, **40**, 243–264. https://doi.org/10.1080/02643294.2023.2275837

Gómez, R., & **Maye, J.** (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, **7**, 183–206. https://doi.org/10.1207/s15327078in0702_4

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, **13**, 431–436. https://doi.org/10.1111/1467-9280.00476

Hoppe, D. B., **Hendriks, P.**, **Ramscar, M.**, & **van Rij, J.** (2022). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. *Behavior Research Methods*, **54**, 2221–2251. https://doi.org/10.3758/s13428-021-01711-5

Houston-Price, C., & **Nakai, S.** (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, **13**, 341–348. https://doi.org/10.1002/icd.364

Huber, E., **Corrigan, N. M.**, **Yarnykh, V. L.**, **Ferjan Ramírez, N.**, & **Kuhl, P. K.** (2023). Language experience during infancy predicts white matter myelination at age 2 years. *The Journal of Neuroscience*, **43**, 1590–1599. https://doi.org/10.1523/jneurosci.1043-22.2023

Hunter, M. A., & **Ames, E. W.** (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, **5**, 69–95.

Isbilen, E. S., & **Christiansen, M. H.** (2022). Statistical learning of language: A meta-analysis into 25 years of research. *Cognitive Science*, **46**, e13198. https://doi.org/10.1111/cogs.13198

Issard, C., & **Gervain, J.** (2018). Variability of the hemodynamic response in infants: Influence of experimental design and stimulus complexity. *Developmental Cognitive Neuroscience*, **33**, 182–193. https://doi.org/10.1016/j.dcn.2018.01.009

Ji, Y., **van Rij, J.**, & **Taatgen, N.** (2025a). How do children acquire syntactic structures? A computational simulation of learning syntax from input. In *International workshop on the acquisition of (syntactic) complexity at the Interface, March 26–27, 2025. Leuven, Belgium*. KU Leuven.

Ji, Y., **van Rij, J.**, & **Taatgen, N.** (2025b). Skill acquisition from a bottom-up perspective. In *23rd annual meeting of the international conference on cognitive modelling. Ohio, USA*. The Ohio State University.

Kosie, J. E., **Zettersten, M.**, **Abu-Zhaya, R.**, **Amso, D.**, **Babineau, M.**, **Baumgartner H. A.**, **Bazhydai, M.**, **Belia, M.**, **Benavides-Varela, S.**, **Bergmann, C.**, **Berteletti, I.**, **Black, A. K.**, **Borges, P.**, **Borovsky, A.**, **Byers-Heinlein, K.**, **Cabrera, L.**, **Calignano, G.**, **Cao, A.**, **Chijiiwa, H.** … **Lew-Williams, C.** (2023). Manybabies 5: A large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time. https://doi.org/10.31234/osf.io/ck3vd_v1

Kotseruba, I., & **Tsotsos, J. K.** (2018). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, **53**, 17–94. https://doi.org/10.1007/s10462-018-9646-y

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, **5**, 831–843. https://doi.org/10.1038/nrn1533

Laird, J. E., **Lebiere, C.**, & **Rosenbloom, P. S.** (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, **38**, 13–26. https://doi.org/10.1609/aimag.v38i4.2744

Laird, J. E., **Newell, A.**, & **Rosenbloom, P. S.** (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, **33**, 1–64. https://doi.org/10.1016/0004-3702(87)90050-6

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means.* R package version 1.4.8.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk (third edition): Volume I: Transcription format and programs, volume II: The database. *Computational Linguistics* **26**, 657–657. https://doi.org/10.1162/coli.2000.26.4.657

Marcus, G. F., **Vijayan, S.**, **Bandi Rao, S.**, & **Vishton, P. M.** (1999). Rule learning by seven-month-old infants. *Science*, **283**, 77–80. https://doi.org/10.1126/science.283.5398.77

Mareschal, D., & **French, R. M.** (2017). Tracx2: A connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372**, 20160057. https://doi.org/10.1098/rstb.2016.0057

Menn, K. H., **Männel, C.**, & **Meyer, L.** (2023). Phonological acquisition depends on the timing of speech sounds: Deconvolution EEG modeling across the first five years. *Science Advances*, **9**, eadh2560. https://doi.org/10.1126/sciadv.adh2560

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual information processing* (pp. 283–308). Elsevier. https://doi.org/10.1016/b978-0-12-170150-5.50012-3

Perruchet, P., & **Vinter, A.** (1998). Parser: A model for word segmentation. *Journal of Memory and Language*, **39**, 246–263. https://doi.org/10.1006/jmla.1998.2576

Pinker, S. (1999). Out of the minds of babes. *Science*, **283**, 40–41. https://doi.org/10.1126/science.283.5398.40

Plante, E., & **Gómez, R. L.** (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools*, **49**, 710–722. https://doi.org/10.1044/2018_lshss-stlt1-17-0131

Rabagliati, H., **Ferguson, B.**, & **Lew-Williams, C.** (2018). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, **22**, e12704. https://doi.org/10.1111/desc.12704

Raz, G., **Cao, A.**, **Bui, M. K.**, **Frank, M. C.**, & **Saxe, R.** (2023). No evidence for familiarity preferences after limited exposure to visual concepts in preschoolers and infants. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **45**, 3319–3325.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, **81**, 149–169. https://doi.org/10.1016/s0010-0277(01)00132-9

Saffran, J. R., **Aslin, R. N.**, & **Newport, E. L.** (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., & **Kirkham, N. Z.** (2018). Infant statistical learning. *Annual Review of Psychology*, **69**, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Simon, H. A., & **Newell, A.** (1971). *Human problem solving: The state of the theory in 1970* (Vol. **26**). American Psychological Association (APA). https://doi.org/10.1037/h0030806

Smith, L., & **Yu, C.** (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, **106**, 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Soderstrom, M. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, **49**, 249–267. https://doi.org/10.1016/s0749-596x(03)00024-x

Soderstrom, M., **Blossom, M.**, **Foygel, R.**, & **Morgan, J. L.** (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, **35**, 869–902. https://doi.org/10.1017/s0305000908008763

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, **120**, 439–471. https://doi.org/10.1037/a0033138

Taatgen, N. A. (2017). Cognitive architectures: Innate or learned? *AAAI fall symposium series*, Technical report FS-17-05.

Taatgen, N. A. (2023). Prims tutorial [GitHub repository]. Retrieved from https://github.com/ntaatgen/PRIMs-Tutorial

Taatgen, N. A., & **Anderson, J. R.** (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition*, **86**, 123–155. https://doi.org/10.1016/s0010-0277(02)00176-2

Taatgen, N. A., **van Vugt, M. K.**, **Daamen, J.**, **Katidioti, I.**, **Huijser, S.**, & **Borst, J. P.** (2021). The resource-availability model of distraction and mind-wandering. *Cognitive Systems Research*, **68**, 84–104. https://doi.org/10.1016/j.cogsys.2021.03.001

Thiessen, E. D., & **Pavlik, P. I.** (2012). iminerva: A mathematical model of distributional statistical learning. *Cognitive Science*, **37**, 310–343. https://doi.org/10.1111/cogs.12011

Ullman, M. T., **Clark, G. M.**, **Pullman, M. Y.**, **Lovelett, J. T.**, **Pierpont, E. I.**, **Jiang, X.**, & **Turkeltaub, P. E.** (2024). The neuroanatomy of developmental language disorder: A systematic review and meta-analysis. *Nature Human Behaviour*, **8**, 962–975. https://doi.org/10.1038/s41562-024-01843-6

Ullman, M. T., **Earle, F. S.**, **Walenski, M.**, & **Janacsek, K.** (2020). The neurocognition of developmental disorders of language. *Annual Review of Psychology*, **71**, 389–417. https://doi.org/10.1146/annurev-psych-122216-011555

van der Velde, M., **Sense, F.**, **Borst, J. P.**, & **van Rijn, H.** (2022). Explaining forgetting at different timescales requires a time-variant forgetting function. Preprint available at https://doi.org/10.31234/osf.io/d58n4

van Rij, J., **van Rijn, H.**, & **Hendriks, P.** (2010). Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language*, **37**, 731–766. https://doi.org/10.1017/s0305000909990560

Visser, I., **Geambasu, A.**, **Baumgartner, H. A.**, **Bergmann, C.**, **Byers-Heinlein, K.**, **Carstensen, C. A.**, **Doyle, F. L.**, **Gervain, J.**, **Hannon, E.**, **Havron, N.**, **Johnson, S.**, **Kachergis, G.**, **Kline Struhl, M.**, **Kosie, J. E.**, **Lew-Williams, C.**, **Mayor, J.**, **Moreau, D.**, **Mueller, J. L.**, **Raijmakers, M. E. J.**, **Shukla, M.**, **Tsui, A.**, **Sirois, S.**, **Westermann, G.**, **Soderstrom, M.**, & **Levelt, C.** (2021) Many babies 3 rule learning – Stage 1 registered report – In principle acceptance. https://doi.org/10.31234/osf.io/aex7v

Walker, N., **Monaghan, P.**, **Schoetensack, C.**, & **Rebuschat, P.** (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, **70**, 221–254. https://doi.org/10.1111/lang.12395

Wilson, B., **Spierings, M.**, **Ravignani, A.**, **Mueller, J. L.**, **Mintz, T. H.**, **Wijnen, F.**, **van der Kant, A.**, **Smith, K.**, & **Rey, A.** (2018). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science*, **12**, 843–858. https://doi.org/10.1111/tops.12381

Cite this article: Ji, Y., van Rij, J., & Taatgen, N. (2025). Simulating procedural discovery in early language acquisition: Domain-general cognition with contextual learning. *Journal of Child Language* 1–40, https://doi.org/10.1017/S0305000925100159