The Role of Structural Input Features in Statistical Learning

Danaja Rutar (rutar.danaja@gmail.com)
Department of Cognitive Psychology and Department of AI

Erwin de Wolff (erwin_de_wolff@live.nl)
Department of AI

Johan Kwisthout (johan.kwisthout@donders.ru.nl)
Department of AI, Thomas Van Aquinostraat 4
6525 GD Nijmegen, The Netherlands

Abstract

Two learning mechanisms have been suggested to underlie statistical learning: computation of transitional probabilities and chunking. It remains an open question though what determines which mechanism is used. In this study, we examined whether learning mechanisms are exploited differentially depending on the structure of the input to be learned. More specifically, we investigated whether the strength of the relationships between elements in the input structure and the presence of higher-order relationships influence the employment of the mechanisms. Participants were presented with three different input structures. We measured reaction times in a self-paced statistical learning task and created Bayesian models that formalised different learning mechanisms. The results show that the employment of the learning mechanisms indeed depends on the input structure. Further studies will need to examine a more specific mapping between the input structures and the learning mechanisms.

Keywords: statistical learning; structural input features; chunking model; transitional probabilities model; hybrid model

Introduction

Humans are very skilled at navigating different environments, and to do so, they need to be able to learn quickly about the structure of their environment. A seminal study demonstrated that humans can indeed detect statistical patterns in a continuous stream of sensory input after only a relatively short period of exposure and in the absence of any instruction or feedback (Saffran, Aslin, & Newport, 1996). Since then, the ability for statistical learning has been thoroughly examined and it has been shown to exist for different sensory modalities and to be present from early in life (Hunnius, 2022; Perruchet, 2019; Saffran & Kirkham, 2018).

Computing transitional probabilities and chunking as learning mechanisms

Different mechanisms have been proposed to underlie statistical learning. One idea is that statistical structure in the environment is extracted by forming pairwise associations between elements in the input and representing their transitional probabilities (Fiser & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Perruchet, 2019). Transitional probability (TP) models learn statistical relations between adjacent elements in a sensory stream and can hence predict an upcoming element based on the preceding element in a sequence. Another mechanism that has been proposed to underlie statistical learning is chunking. Chunking extracts frequently occurring, statistically-coherent units from the input without computing TPs between their elements (Fiser, 2009; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet, 2019; Slone & Johnson, 2018). It assumes economy of representation because sub-units are progressively forgotten as bigger chunks are learned.

Different factors influence the employment of learning mechanisms

Twenty years after the seminal study on statistical learning (Saffran et al., 1996), it remains unclear whether humans use one learning mechanism or whether they flexibly use several mechanisms. Various factors have been proposed to influence the employment of learning mechanisms and the nature of statistical learning, such as availability of *temporal cues* in sequences – which promote chunking; *dips* in distribution of the sequences – which are best detected by TP models (Franco & Destrebecqz, 2012); *time of exposure* – with TP models being better predictors of initial learning outcomes and chunking models being better predictors of learning after longer exposure to the statistical patterns (Slone & Johnson, 2018); and *input modality* – with hippocampus, basal ganglia and thalamus being involved across all modalities (Batterink, Paller, & Reber, 2019; Frost, Armstrong, Siegelman, & Christiansen, 2015) and certain brain areas being modality specific (Conway, 2020; Frost et al., 2015). A particularly relevant factor whose effect on the deployment of the learning mechanisms has been rarely investigated is the structure of the sensory input itself.

Input structure can vary in several ways, including the number of preceding elements needed to predict an upcoming element (Gomez, 1997), the strength of a relationship between two elements (probabilistic vs. deterministic) (Wilson et al., 2013), the size of sequences to be learned, the presence of higher-order patterns (Fiser & Aslin, 2001; Fitch & Friederici, 2012; Fitch & Martins, 2014) and non-adjacent relationships (Gomez, 2002; Remillard, 2008). These structural features could be mapped on a continuum, with deterministic, adjacent, and linear dependencies between elements at the one end, and probabilistic, non-adjacent, and higher-order dependencies at the other (Conway, 2020). One end of the continuum is mediated by implicit processing and proceeds automatically, engaging posterior brain regions. The other end is mediated by more explicit processing and demands more attention, engaging frontal brain regions (Con-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

way, 2020; Fuster & Bressler, 2012). The above reasoning was taken by some as possibly suggesting that humans have different learning mechanisms to learn different input structures (Bahlmann, Gunter, & Friederici, 2006; Conway, 2020; Conway & Christiansen, 2001; Uddén & Bahlmann, 2012). Additionally, there are certain kinds of higher-order patterns in the environment such as conjunctive relationships that are not learned well by chunking and TP models yet both adults and children can learn them (Lucas, Bridgers, Griffiths, & Gopnik, 2014). Hence, additional learning mechanisms might exist – "hybrid mechanisms" – that are grounded in existing TP and chunking mechanisms but combine features of the two in ways that allow them to learn higher-order relation such as conjunction.

The effect of input structure on the employment of learning mechanisms

These studies provide first valuable insights into the differences in input structures and their possible impact on statistical learning. Building on this, we investigated whether humans utilise other statistical learning mechanisms than TP learning and chunking and whether the structure of the input determines which learning mechanism is employed. Specifically, we explored whether the strength of the relationships between elements in the input structure (i.e., more probabilistic vs. more deterministic) and the presence of higher-order relationships (in particular conjunctive relationship) determine the mechanism used by the learner. We created three 3-element input structures placed at different positions along this continuum (see Table 1 and Figure 1b). For example, Input structure 1 consisted of only probabilistic relationships between its elements and required learning a higher-order relationship. This relationship stipulated that only the combination of the first two elements could predict the third one. On the other hand, Input structure 2 was situated at the opposite end of the complexity continuum. It contained a higher-order relationship that could be learned but was not necessary to predict the final element and it included one deterministic relationship. Finally, Input structure 3 occupied an intermediate position between these two extremes, with only probabilistic relationships and no requirement to learn a higher-order relationship.

Table 1: The three input structures and their characteristics.

Input Structure	**Probabilistic relationships**	**Higher-order relationships**
2	+	- (present, learning not required)
3	++	-
1	++	+

Bayesian models as implementations of learning mechanisms

To investigate the effect of the input structures on the employment of learning mechanisms, we devised an experimental self-paced task (based on (Siegelman, Bogaerts, Kronenfeld, & Frost, 2018)), measured reaction times (RTs) (which we assumed reflected prediction error), created Bayesian models that formalised different learning mechanisms, and compared the measured RTs with the models' predictions to determine which model best described learning of each input structure.

As we wanted to explore a wide range of structural relations that have not been necessarily well captured by the traditional TP and chunking models we complemented the existing chunking and TP models with three additional learning models (see Figure 2) which are either enhancements or combinations of the two existing models.

Materials and Methods

Participants

Forty-eight healthy adult participants took part in our study and forty-four were included in the final analysis (14 males, 30 females, mean age = 28.0 years, $SD = 9.2$). One participant was excluded for not following instructions, and three were excluded for taking significantly longer to complete the task.

The Ethics Committee of the Faculty of Social Sciences (ECSS) of Radboud University Nijmegen, The Netherlands, approved the study (approval number ECSW2016-0905-396) and all participants gave written informed consent. Participants received either a monetary compensation or course credit for participating in the experiment.

Materials, task, and procedure

Three input structures were created from 21 different shapes, each with a unique colour (see Figure 1a). The input structures differed in the strength of the relationships between elements in the input structure (i.e., more probabilistic vs. more deterministic) and in whether they required learning of higher-order conjunctive regularities (see Figure 1b). There were four versions of each input structure, and each consisted of different shapes. Shapes for each version of the input structure were constant within participants but varied between participants. During the task, there were 22 repetitions of each input structure version in a random order. All participants were exposed to all three input structures. The self-paced task (Siegelman et al., 2018; Siegelman, Bogaerts, Armstrong, & Frost, 2019) was structured such that the shapes would appear on the screen one after another with one shape at a time being present on the screen (Figure 1c). To advance from one shape to the next, participants needed to press the letter 'K'. RTs of each 'K' press after the stimulus presentation were recorded and used as input for the Bayesian models in the main analysis.

Before the task, participants were told that they would be shown a sequence of shapes, that some shapes tended to follow one another and that their task was to find patterns. Pat-

terns could be of different lengths and could be structured in different ways. They were also told that the experiment would take between 20 and 30 minutes, after which they would be asked to report any patterns they observed during the study.

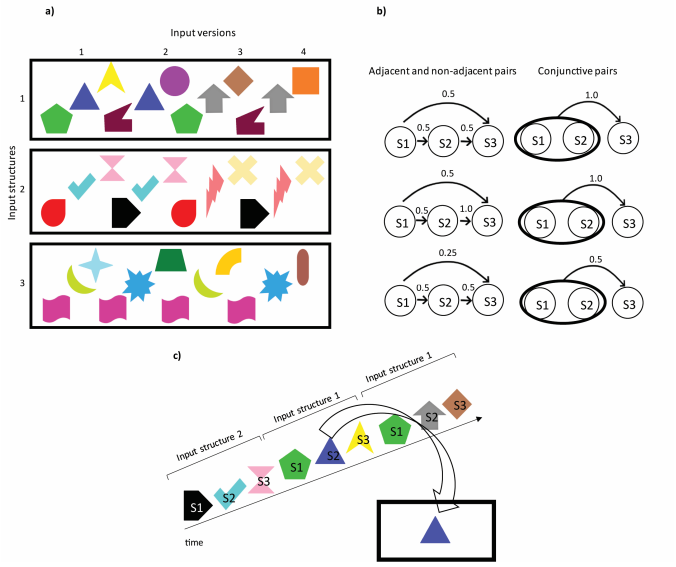

Figure 1: Task materials and task presentation. **(a)** Input structures and input versions. **(b)** Statistical patterns underlying the three input structures. Each input structure consisted of three shapes (i.e., *S1*, *S2*, *S3*) and differed in TPs between the shapes. On the left are TPs between adjacent and non-adjacent pairs and on the right is a TP between the conjunctive pair and the third shape. Left and right representation are both about the same input structure, they just focus on different aspects represented in individual input structure. **(c)** Task presentation. Input structures followed one after another in a random order (1c, above) and participants saw one imagine at a time on the screen (1c, below).

Formal Bayesian models

We constructed and implemented five Bayesian models that represent five different learning mechanisms (see Figure 2). We wanted to explore a wide range of structural relations that have not been necessarily well captured by the traditional TP and chunking models, therefore we complemented the existing chunking and TP models with three additional learning models which were either enhancements or combinations of the two existing models.

Chunking model: The chunking model learns the distribution over a compound variable $P(S1S2S3)$. This model is distinct from other models because updating of the distribution occurs only after shapes $S1, S2$ and $S3$ have been seen.

TP model: This model learns the probability of *Sm*, $P(Sm)$, and the probability of *Sn* given *Sm*, $P(Sn|Sm)$. Subscripts m and n denote two consecutive shape positions in a stream of shapes.

TP-connected model: This model differs from the chunking model, as it represents three separate probability distributions: the probability of $S1$, $P(S1)$, the probability of $S2$ given $S1$, $P(S2|S1)$, and the probability of $S3$ given $S1$ and $S2$, $P(S3|S1,S2)$. As a consequence, the TP-connected model learns both, adjacent and non-adjacent TPs and thus can be seen as an enhanced version of the traditional TP model.

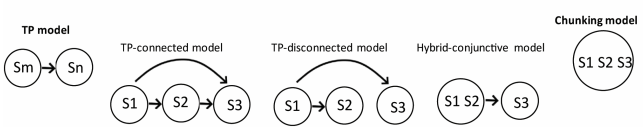

Figure 2: Graphical representation of formal Bayesian models. At the two ends of the continuum are (bolded) the two main models: the TP model, which learns the dependency between *Sm* and *Sn*, and the chunking model, which learns the joint probabilities for each combination of *S1*, *S2,* and *S3*. In between the two main models are enhanced TP models (TP-connected and TP-disconnected model) and the hybrid model that is a combination of the chunking and the TP model. The TP-connected model learns the dependency between *S1* and *S2* and between *S2* and *S3* as well as the dependency between *S1* and *S3*. The TP-disconnected model learns the dependency between *S1* and *S2* and the dependency between *S1* and *S3,* but not the dependency between *S2* and *S3*. The hybrid-conjunctive model learns the joint probability of *S1* and *S2* and the relation between the combined *S1*, *S2* and *S3*.

TP-disconnected model: The TP-disconnected model is another enhanced version of the TP model. Like the TP-connected model, the TP-disconnected model learns the probability of $S1$, $P(S1)$, and probability of $S2$ given $S1$, $P(S2|S1)$. However, in the TP-disconnected model $S3$ is only predicted based on $S1$, but not $S2$, so $P(S3|S1)$. As such, the TP-disconnected model is more parsimonious when it comes to predicting the final element of the input structure, but it omits the last connection.

Hybrid-conjunctive model: The conjunctive model learns the distribution over a compound variable $P(S1S2)$, and the probability of $S3$ given the distribution over a compound variable of $S1$ and $S2$, $P(S3|S1S2)$. As the combined-conjunctive model learns the first and the second element as a chunk and then learns the connection between the second and the third element as a TP, it is a hybrid of chunking and TP models.

Normative modelling – Bayesian models learning the input structures

We formalised learning as the Bayesian updating of Dirichlet distributions (Castillo, Hadi, & Solares, 1997). In our case, the categorical distribution defined the probability distribution over the possible shapes. For each separate conditional probability distribution within the models there was a corresponding Dirichlet distribution. Prior to any evidence, each possible distribution over the shapes was equally likely.

Normative modelling – Relationship between models' prediction errors and RTs

Each Bayesian model processed the observations of the current shapes and generated predictions of the upcoming shapes, where both predictions and observations were probability distributions over all the possible shapes. As the models did not generate direct information on RTs, we derived predicted RTs from what the models predicted and what was observed as follows.

It is well known that RTs are slower after erroneous predictions than after correct predictions (Steinborn, Flehmig, Bratzke, & Schröter, 2012). This effect is known as the post-error slowing effect (Laming, 1979) and is particularly large in self-paced tasks (Jentzsch & Leuthold, 2006). Erroneous predictions, resulting from the divergence between the predicted and observed stimulus, can be cast as prediction error (e.g., (Friston, 2009)). We modelled prediction error as Kullback-Leibler Divergence which measures how much one probability distribution Q is different from another probability distribution P and can be defined as follows (e.g., (Géron, 2022)):

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right)$$

where P is the observation and Q is the (stochastic) prediction made by the model.

Statistical modeling – Model fitting

Prior to fitting the participant RTs to the models, the data was cleaned. We removed RTs higher than 3 SD from the participant's mean and lower than 100 ms (cf. (Dahan, Bennet, & Reiner, 2019; Siegelman et al., 2019)). Then, RTs were log-transformed and normalised using a z-transformation.

Note that in addition to Bayesian models, we also fitted a baseline model to the participants' data. The baseline model always predicts a uniform distribution across all possible shapes. This model does not learn and as such will always experience the exact same surprise from any observed shape. The inclusion of the baseline model allowed us to better asses the absolute rather than relative fit of the Bayesian models.

The goodness of fit was determined based on the posterior probability of each model given the input structure. Because we had no reason to assume that one model would be best, we defined a uniform prior probability distribution $P(M)$ over our normative models. To calculate the posterior probabilities for each model, we defined a likelihood function of the participant's RTs given the predicted RTs generated by a model M. To that end we first modelled the participant's RT given a single shape as the predicted RT plus independent noise. We assumed that this noise was drawn from a normal distribution with 0 as the mean and a standard deviation of 1. Thus, a participant's RT was defined as:

$$y_t = x_t + \mathcal{N}(0,1)$$

where x_t represented the predicted RT and y_t the participant's RT. Since adding a constant to a normal distribution is equal to adding that constant to the mean of the normal distribution, this equation is equivalent to:

$$y_t = \mathcal{N}(x_t,1)$$

To find the likelihood of a particular participant's RT given the predicted RT, we looked at the density height of this normal distribution at the participant's RT:

$$\mathcal{L}(y_t|x_t) = \mathcal{N}(y_t|x_t,1)$$

The likelihood of the entire series of the participant RTs given the predicted RTs is given by:

$$\mathcal{L}(y|x) = \prod_t \mathcal{L}(y_t|x_t)$$

To determine the model that best explains participant RTs, we equate a particular model $m \in M$ with the predicted RTs $\mathbf{x}_m$ it generates, such that $m = \mathbf{x}_m$. We calculated the posterior probability over the models by applying Bayes' rule:

$$P(M|y) \propto \mathcal{L}(y|M) \times P(M)$$

This provides us with the posterior probability over different models for a single participant. Since we were interested in the average posterior probability of the models across participants, we averaged the individual posterior probabilities for each model. To determine which of the models was the best fit for a particular input structure, we calculated the posterior for each input structure separately.

Results

Group level behavioural findings

To assess whether participants learned to predict the shapes over time, we tested for decreases in reaction times (RTs) across trials. We computed Spearman correlations between log-transformed RTs and trial number for each input structure, finding significant negative correlations: Input 1 (r = –0.76), Input 2 (r = –0.75), and Input 3 (r = –0.73), all $p < .001$, $n = 265$. This indicates that participants on average became faster as the experiment progressed.

Group level modelling findings – model preference at the end of learning

Based on the posterior probabilities of models in Table 2, we can conclude that the baseline model was most frequently employed for learning Input structure 1, $P(M_{baseline}|I_1) = 0.51$, closely followed by the TP-disconnected model $P(M_{TP-disconnected}|I_1) = 0.45$. The TP-disconnected model was most frequently employed for learning Input structure 2, $P(M_{TP-disconnected}|I_2) = 0.58$, and the TP-connected model was most often employed for learning Input structure 3, $P(M_{TP-connected}|I_3) = 0.70$. Our hypothesis that different input structures should be learned using different learning mechanisms was mostly supported, as our participants displayed RT patterns that fitted best with two different models, the TP-disconnected and the TP-connected model.

Group level modelling findings – model preference throughout learning

To gain further insight into how evidence for models was changing over time we performed an additional, trial-by-trial analysis (see Figure 3). We can interpret trial-by-trial posterior probabilities as showing how much evidence there was for each model at each moment in time on a group basis. Posterior probabilities of all models during the first part of the task were similar, indicating no particular preference for any model. Then, a clear preference for one model began to form for all three input structures, and in the last part the posterior probability of the baseline model increased and approached (and even exceeded in the case of Input structure 1) the posterior probability of the model that had the highest posterior probability up to that point.

Individual level modelling findings

We conducted an additional analysis to examine whether participants differed in their preferred learning mechanisms (see Figure 4). Evidence for individual differences would be present if either (a) each participant consistently relied on a single model across all input structures, with different participants favouring different models, or (b) each participant used a distinct combination of models across input structures, suggesting individualized strategies. Overall, the results provide only limited evidence for individual differences in preferred learning mechanisms. No participants consistently favoured a single model across input structures (in contrast to a), and the majority showed a similar pattern of model usage, with several models – particularly the TP-disconnected and TP-connected models – frequently dominating across participants (in contrast to (b)). This pattern indicates a small degree of idiosyncrasy, but it is not pronounced or systematic enough to support strong claims about stable individual differences in learning strategies.

Discussion and Conclusions

The aim of our study was to investigate whether input structure plays a role in determining which learning mechanism is employed. It has been suggested that different structural features could be mapped on a continuum, with deterministic, adjacent, and linear dependencies between elements at the one end, and probabilistic, non-adjacent, and higher-order dependencies at the other. Different parts of the continuum have different cognitive requirements and could also be mediated by different neural processes, indicating that at least two distinct learning processes may be involved in statistical learning (Bahlmann et al., 2006; Conway, 2020).

The results of our study support the hypothesis that the structure of the input influences the employment of the learning mechanism which is consistent with previous suggestions that humans might have different learning mechanisms to learn different input structures (Bahlmann et al., 2006; Conway, 2020; Conway & Christiansen, 2001; Uddén & Bahlmann, 2012). Below, we discuss how learning of each input structure might have happened.

For learning the Input structure 1 participants were using the TP-disconnected model until the very end and in the last few trials the baseline model took over. Whereas evidence exists that adults, and even children, are able to recognise conjunctive relationships where two events together (but not separately) lead to a third event (Lucas et al., 2014), the fact that the participants' data is best described by the baseline and the TP-disconnected model suggests that they did not learn the conjunctive relationship in the structure. The employment of the TP-disconnected model implies that the participants learned that there was a connection between the first and the last element, but failed to learn that the first and second element *jointly* predicted the last element. Ultimately, the baseline model was the best predictor of participants' RTs, which indicates that, in the last few trials, participants were not learning anymore, probably due to fatigue and boredom. Although unexpected, this finding can be explained by the fact that the Input structure 1 was the most complex, as it contained only probabilistic relationships and a higher-order relationship.

Participants' RTs when learning Input structure 2 were best described by the TP-disconnected model. It is possible that the participants first learned the relationship between all shapes in this input structure as predicted (thus employing the TP-connected model, which was predominant model between trials 20 and 35 as seen in Figure 3) before wrongly removing the connection between the second and the third element (thus using the TP-disconnected model).

Input structure 3 was best learned by the TP-connected model. That is, participants have learnt not only the relations between the adjacent elements (TP model) but also the relations between the first and the last element (TP-connected model). This is an unusual finding given that the TPs between all elements, adjacent and non-adjacent were, were weak.

Importantly, participants were exposed to all three input structures in a single, fully interleaved stream. Although interleaving can, in principle, create order effects and recruit mechanisms that differ from blocked learning (Zhou, Singh, Tandoc, & Schapiro, 2023), three features of the present design mitigate this concern. First, item order was randomised for each participant, eliminating systematic sequence confounds. Second, RT analyses and Bayesian model fits were computed separately for each structure – any cross-talk would have emerged as structure-specific asymmetries, which were not observed. Third, learning signatures (RT declines and dominant model posteriors) were highly consistent across structures, showing that prior exposure neither facilitated nor impeded subsequent learning. Because interleaving better approximates real-world input streams, we view it as a strength rather than a limitation. Nevertheless, future work could contrast blocked and interleaved presentation to further isolate potential order effects.

A general trend that we observe across all input structures is that the connection between the first and second and the

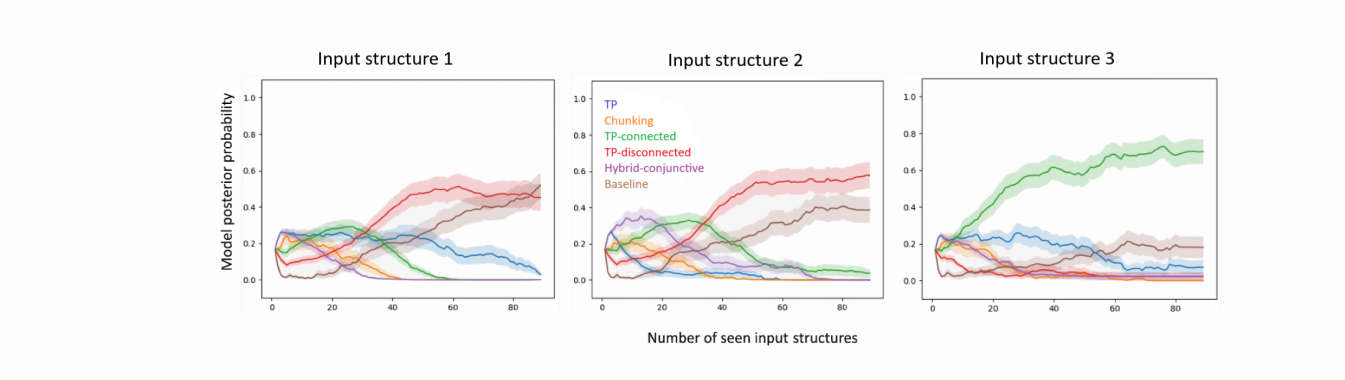

Figure 3: Trial by-trial posterior probabilities of different models for the three input structures. TP model: blue, chunking model: orange, TP-connected model: green, TP-disconnected model: red, hybrid-conjunctive model: purple, baseline model: brown.

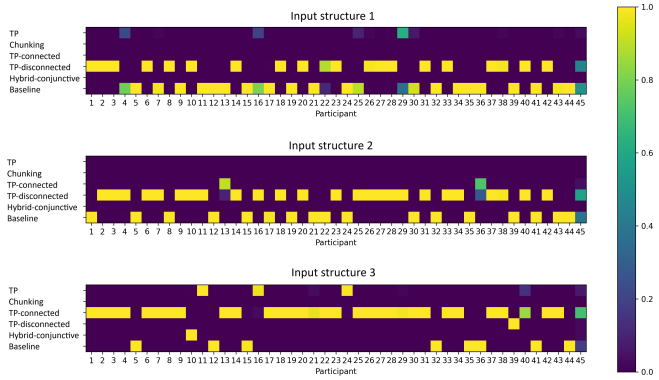

Figure 4: Individual level posterior probabilities of different models for the three input structures. Yellow indicates posterior probability of 1.0, i.e., maximal model evidence, and dark blue indicates posterior probability of 0.0, i.e., minimal model evidence.

first and last element was always established, anchoring all subsequent predictions of elements to the first element. This results in an economic structural representation because it allows participants to represent all elements in one input structure and to infer the length of an input structure. This might further suggest that participants prefer to see individual elements as belonging to bigger input structures but individual relations between the subsequent elements in each input structure might not be always important to capture. Unfortunately, our study cannot sufficiently answer *why* some but not other relations were represented and why different relations were represented across the input structures. One way to approach this question would be to investigate utility functions that evaluate the *efficacy* of representing different relations, such as for example trading off between accuracy and complexity (Jefferys & Berger, 1992).

Additionally, our study also does not provide answers as to *how* structural input features affect the learning mechanisms. Several experimental approaches could help elucidate this relationship. First, the exact mapping between learning mechanisms and structural input features would need to be established by manipulating individual structural input features. In our study, we combined several features into one input structure, whereas future studies should create input structures that vary only one input feature at a time. For example, separate experiments could systematically vary the features investigated in our study: The strength of relationships (from fully deterministic to increasingly probabilistic), the adjacency of dependencies (from adjacent to progressively more distant elements) and the complexity of patterns (from linear to various types of hierarchical structures). Furthermore, computational modeling approaches could help identify additional features of input structures that might drive the selection of specific learning mechanisms. This could involve developing formal metrics to quantify structural complexity, information content, and pattern regularity. Such metrics could then be used to predict when learners might switch between different learning strategies.

Whilst many questions remain about how exactly structural features recruit distinct learning mechanisms, the added value of our study is in taking the first step toward operationalising these theoretical distinctions and introducing a novel modeling framework to investigate them.

Acknowledgements

This work was supported by the Donders Centre for Cognition, Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, through the internal grant "Understanding predictive processing in development: Modelling the generation of generative models." We would also like to thank the anonymous reviewers for their constructive and helpful comments.

References

Bahlmann, J., Gunter, T. C., & Friederici, A. D. (2006). Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience*, *18*(11), 1829–1842.

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the neural bases of implicit and statistical learning. *Topics in Cognitive Science*, *11*(3), 482–503.

Castillo, E., Hadi, A. S., & Solares, C. (1997). Learning and updating of uncertainty in dirichlet models. *Machine Learning*, *26*, 43–63.

Conway, C. M. (2020). How does the brain learn environmental structure? ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience & Biobehavioral Reviews*, *112*, 279–299.

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*(12), 539–546.

Dahan, A., Bennet, R., & Reiner, M. (2019). How long is too long: An individual time-window for motor planning. *Frontiers in Human Neuroscience*, *13*, 238.

Fiser, J. (2009). Perceptual learning and representational learning in humans and animals. *Learning & Behavior*, *37*(2), 141–153.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1933–1955.

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*(1), 87–104.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125.

Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow*. O'Reilly Media, Inc.

Gomez, R. L. (1997). Transfer and complexity in artificial grammar learning. *Cognitive Psychology*, *33*(2), 154–207.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Hunnius, S. (2022). Early cognitive development: Five lessons from infant learning. In *Oxford research encyclopedia of psychology.* Oxford University Press.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist*, *80*(1), 64–72.

Jentzsch, I., & Leuthold, H. (2006). Short article: Control over speeded actions: A common processing locus for micro-and macro-trade-offs? *Quarterly Journal of Experimental Psychology*, *59*(8), 1329–1337.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, *43*(3), 199–224.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? transitional probabilities versus chunks in language learning. *Topics in Cognitive Science*, *11*(3), 520–535.

Remillard, G. (2008). Implicit learning of second-, third-, and fourth-order adjacent and nonadjacent sequential dependencies. *Quarterly Journal of Experimental Psychology*, *61*(3), 400–424.

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

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

Siegelman, N., Bogaerts, L., Armstrong, B. C., & Frost, R. (2019). What exactly is learned in visual statistical learning? insights from bayesian modeling. *Cognition*, *192*, 104002.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, *42*, 692–727.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition*, *178*, 92–102.

Steinborn, M. B., Flehmig, H. C., Bratzke, D., & Schröter, H. (2012). Error reactivity in self-paced performance: Highly-accurate individuals exhibit largest post-error slowing. *The Quarterly Journal of Experimental Psychology*, *65*(4), 624–631.

Uddén, J., & Bahlmann, J. (2012). A rostro-caudal gradient of structured sequence processing in the left inferior frontal gyrus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2023–2032.

Wilson, B., Slater, H., Kikuchi, Y., Milne, A. E., Marslen-Wilson, W. D., Smith, K., & Petkov, C. I. (2013). Auditory artificial grammar learning in macaque and marmoset monkeys. *Journal of Neuroscience*, *33*(48), 18825–18835.

Zhou, Z., Singh, D., Tandoc, M. C., & Schapiro, A. C. (2023). Building integrated representations through in-

terleaved learning. *Journal of Experimental Psychology: General*, *152*(9), 2666.