Timesteps of Mamba Align with Human Reading Times

Yuji Yamamoto[1,2*] **Shinnosuke Isono**[2*] **Yoshinobu Kawahara**[3,5] **Sho Yokoi**[2,4,5]

[1]SOKENDAI [2]NINJAL [3]The University of Osaka [4]Tohoku University [5]RIKEN
{yuji.yamamoto, s-isono, yokoi}@ninjal.ac.jp kawahara@ist.osaka-u.ac.jp

Abstract

This study demonstrates an alignment of per-word processing time in a popular state-space language model Mamba and human readers. In Mamba, the recurrent state transition at each layer conceptually takes some duration of time, the discretization timestep Δ_t, determined dynamically in response to the input. Using a naturalistic reading dataset, we show that the per-word timestep from Mamba is a significant predictor of human reading times, and remains significant even when known predictors such as GPT-2 surprisal are controlled for. We further suggest, through formal analysis of Mamba's architecture and internal dynamics, that Mamba can serve as a new, valuable lens to look at human real-time language processing with ever-updated memory, because it allows us to look at how each module (layer) weighs short- and long-term information retention, and how noise may interact with dynamic, continuous memory representation. Code is available online.[1]

1 Introduction

Human language inherently unfolds over time. In the rise of fluent artificial large language models (LMs), it has become interesting to ask how time may be represented in (unidirectional variants of) them, and how their time is related to temporal dynamics of language processing in humans.

Mamba (Gu and Dao, 2024), a popular architecture among LMs based on a state-space model (SSM), is interesting in this respect, since it has a notion of internal timesteps, which modulates state transitions in response to input. SSMs represent sequential data as continuous state transitions and are characterized by their ability to handle long-range dependencies in a stable manner. Among them, Mamba is designed as a selective SSM that

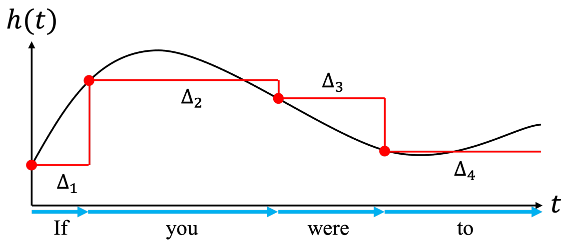

Figure 1: Illustration of discretization of a continuous-time state transition $h(t)$ using the input-dependent timestep Δ_t.

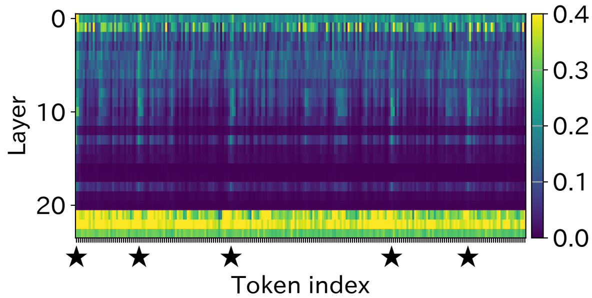

Figure 2: Discretization timesteps $\bar{\Delta}_t$ of each of the 24 layers of Mamba-130M, in response to the first five sentences from the Natural Stories Corpus. "★" indicates the beginning of a sentence.

dynamically modulates state transitions by changing the size of timesteps in response to the input, offering greater flexibility than other SSMs that assume fixed transition dynamics. As illustrated in Figure 1, this property enables the model to immediately reflect changes of context and importance of particular input in its internal states, allowing adaptive representation learning even for data with complex and non-stationary structures.

In this study, we demonstrate a word-level alignment between Mamba's internal timesteps and human processing time. After establishing that Mamba's internal timesteps can be seen as processing time (§3), we show that the sizes of the timesteps that Mamba assigns to each word in a

*Equal contribution
[1]Code available at https://osf.io/vnw5e/overview?view_only=93ad704fc6ea44438f3d3538b4b682eb

Findings of the Association for Computational Linguistics: ACL 2026, pages 31820–31832

naturalistic corpus is predictive of the time that human readers spend on each word (§4). The performance of the best-performing layer is comparable to that of GPT-2 surprisal, a variable known to be a powerful predictor of human reading times, and remains significant even when GPT-2 surprisal and other known predictors are controlled for. We also conducted formal analyses looking into how Mamba's architecture and internal dynamics process language over time, and argue that the model can serve as a new, valuable lens to look into human real-time language processing with ever-updated memory (§5). Specifically, we observe that the model's transition matrix eigenspectrum specifies a different degree of memory retention ability for each layer, enabling the entire model to process both short- and long-term information. We also theoretically show that the size of the timesteps corresponds to uncertainty in state transition, and suggest that this offers a new insight into noisy language processing.

2 Preliminaries

2.1 Mamba

We provide an overview of Mamba; for further details, see (Gu and Dao, 2024).

Mamba is a popular state-space model (SSM)-based language model. Unlike Transformers, SSMs do not process the entire sequence simultaneously but instead handle input and output recursively and sequentially on a word-by-word basis, in a similar fashion to RNNs and human readers. Because SSMs are recursive, a naive implementation would require computation time proportional to the input sequence length. However, since the hidden-state update equations of SSMs contain no nonlinearities, the model can be accelerated using parallel scan algorithms (Blelloch, 1990), making large-scale language model training feasible. Mamba, one of the representative state-space models, as well as Linear Attention (Yang and Zhang, 2024) models that also employ linear state updates, are employed in the development of highly efficient large language models (Zuo et al., 2024; Preferred Networks et al., 2025; Kimi Team et al., 2025; Blakeman et al., 2025).

Mamba has multiple layers (for example, the Mamba-130M model used in the current experiment consists of 24 layers[2]). Each layer incorporates a discretized form of a state-space model. While state-space models are defined as differential equations in continuous time, they are approximated as difference equations in discrete time when applied to real data, particularly discrete data such as language (Figure 1). In standard discretization procedures, a constant discretization timestep Δ is used regardless of the time index t; in contrast, Mamba assigns an *input-dependent discretization timestep* Δ_t to each timestep t.

The discretization timestep Δ_t acts as a gate in each layer of Mamba, governing memory retention and forgetting. Below, we formally explain this by referring to the specific update equations of Mamba. Given an input $\boldsymbol{x}_t \in \mathbb{R}^d$, Mamba applies the following equations independently to each element x_{ti} of $\boldsymbol{x}_t$ in parallel d times, and outputs the vector $\boldsymbol{y}_t \in \mathbb{R}^d$ obtained by collecting the resulting outputs y_{ti}. Mamba consists of two equations: one updates a component called the hidden state $\boldsymbol{h}_t \in \mathbb{R}^n$, which corresponds to memory, using the input x_{ti}, and the other generates the output y_{ti} from the hidden state $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \overline{A}_t \boldsymbol{h}_{t-1} + \overline{B}_t x_{ti} \tag{1}$$

$$y_{ti} = C_t \boldsymbol{h}_t + D x_{ti}. \tag{2}$$

Here, the state transition matrix $\overline{A}_t \in \mathbb{R}^{n \times n}$ and input matrix $\overline{B}_t \in \mathbb{R}^{n \times 1}$, which represent the weights for memory and input in the state update equation, are defined as follows with an input-dependent discretization timestep $\Delta_t \in (0, \infty)$:

$$\overline{A}_t = \exp(\Delta_t A) \;\left(= \exp(A)^{\Delta_t}\right) \tag{3}$$

$$\overline{B}_t = \Delta_t W_B \boldsymbol{x}_t \tag{4}$$

$$\Delta_t = \log(1 + \exp(W_\Delta \boldsymbol{x}_t + \boldsymbol{b}_\Delta)). \tag{5}$$

As can be seen from the above equations, the discretization timestep Δ_t plays an important role in modulating the weight coefficients for memory and input. For example, when Δ_t is large, the elements of the matrix $\overline{A}_t$ approach zero, [3] thereby reducing the influence of the previous memory $\boldsymbol{h}_{t-1}$. At the same time, the absolute values of the input weight coefficients $\overline{B}_t$ increase, strengthening the influence of the input $\boldsymbol{x}_t$.[4]

[2] https://huggingface.co/state-spaces/mamba-130m-hf

[3] In Mamba, the transition matrix $\exp(A)$ is implemented to be diagonal, with elements constrained to lie within the range $(0, 1]$.

[4] Conversely, when Δ_t is small, that is, close to zero, the elements of $\overline{A}_t$ approach one and those of $\overline{B}_t$ approach zero, preserving the previous memory $\boldsymbol{h}_{t-1}$ while discarding the input $\boldsymbol{x}_t$.

2.2 Reading Time Modeling

Reading time modeling is a research approach that aims to quantitatively identify factors contributing to cognitive load during sentence comprehension by explaining variations in the reading time that readers spend reading each word in a text. Reading time is considered to reflect comprehension difficulties and information processing complexity, and psycholinguists have attempted to identify factors contributing to cognitive load by statistically analyzing its variation. Many previous studies in this approach use naturalistic reading time datasets, which are collections of reading time data from tens to hundreds of participants reading naturalistic texts (e.g. Demberg and Keller, 2008; Wilcox et al., 2023). The data is modeled by some regression models that contain the variable of theoretical interest plus control variables that are known to affect reading time, such as word length, lexical frequency, syntactic complexity, and contextual predictability. The significance of an explanatory variable suggests that the variable correlates with language processing difficulty, leading to a constructivist understanding of which information is crucial for reading comprehension.

Most studies interested in the relation between large language models and human reading today use the next word probability to link the models and reading times. The *surprisal theory* (Hale, 2001; Levy, 2008) predicts that the processing cost of word w given the preceding context C is proportional to $-\log p(w \mid C)$. The combination of the Transformer architecture and the surprisal theory has proven successful in reading time modeling for naturalistic texts in many languages (Wilcox et al., 2023). Studies have also found, however, that models that are better from an engineering perspective, with more parameters, lower perplexity, or larger context window, are not necessarily better at predicting human reading times than smaller models (Oh and Schuler, 2023; Kuribayashi et al., 2022, 2021). While these studies seek a link between models and humans at the computational level in the sense of Marr (Marr, 1982), other studies seek to link Transformers and humans at the algorithmic level, e.g., comparing the attention mechanism to the memory retrieval mechanism in humans (Ryu and Lewis, 2021; Yoshida et al., 2025).

The current study explores yet another approach to reading time modeling, with a different type of model (Mamba) and a different linking hypothesis (intepreting Δ_t as processing time) than the standard combination of Transformer and surprisal.

3 Discretization Timesteps of Mamba as Word Processing Time

We describe that the discretization timestep Δ_t corresponds to the "processing time" of word t and that this quantity adapts to the input.

Notably, the input-dependent discretization timestep Δ_t (Eq. (5)) in the language model Mamba can be interpreted as the implicit "processing time" involved in state transitions in continuous time. Typically, in the discretization of differential equations, the time evolution in continuous time is approximated using a predetermined fixed step size. In contrast, in Mamba, when modeling continuous-time state transitions as discrete recurrence relations, the time is discretized into timesteps whose length varies depending on the input to each token (Figure 1 and 2). In other words, how long it takes to update the state is controlled by the input tokens.

Mamba potentially captures the processing time allocated to each token through the input-dependent discretization timestep Δ_t. Mamba does not process each token on a fixed time scale; instead, it dynamically adjusts the timesteps Δ_t according to the nature of the input tokens. For example, when the timestep Δ_t is larger, the model continues its time evolution over a longer period while holding the same state. This can be interpreted as a mechanism that adjusts the processing time based on the input. This property is functionally analogous to reading time, where humans pause longer on words or contexts that require a higher cognitive load during comprehension.

4 Mamba's Timesteps are Predictive of Human Reading Times

The current experiment investigates whether Mamba's "processing time", Δ_t, is predictive of human word-by-word reading times, and whether its predictive power is beyond that of other known variables.

4.1 Settings

4.1.1 Datasets

We used two reading time datasets.

Natural Stories Natural Stories (Futrell et al., 2021) consists of data from 181 native speakers of

English reading 10 stories (10,245 words in total) using the self-paced reading method (Just et al., 1982). In this method, one word is presented at a time, and participants press a key to proceed to the next word; no backtracking is allowed. The reading time for a word refers to the duration for which the word is displayed.

OneStop OneStop (Berzak et al., 2025) (the "ordinary reading" subset) consists of reading time data from 180 native speakers of English reading 30 articles using the eye-tracking-during-reading method. The articles come with the original "advanced" version (19,428 words) and the simplified "elementary" version (15,737 words); both were used in the current analysis. We analyze first-pass time (the time between the entry to the region and the first exit to the left or right in the first pass) and regression-path duration (the time between the first entry to the region and the first exit to the right), since they can be seen as conceptual counterparts of Mamba's Δ_t: they reflect the processing time before any information from the right is accessed.

4.1.2 Predictors

Discretization timestep The discretization timestep Δ_t was taken from Mamba-130M and Mamba-2.8B. As described in §2.1, the update equation (1) in Mamba is applied in parallel across d dimensions, so the discretization timestep is obtained as a d-dimensional vector $\mathbf{\Delta}_t$. In the reading-time modeling analysis, we take the average $\bar{\Delta}_t = \sum_i (\mathbf{\Delta}_t)_i$ to aggregate the discretization timestep into one scalar for each layer and each timestep, and it was further aggregated at the word-level by taking the maximal value, when the word was analyzed into multiple subwords.

Control predictors Besides Δ_t, we included the following as control variables: the number of characters in the word, corpus frequency of the word, position of the word in the sentence in the story, Mamba surprisal, and GPT-2 surprisal.[5]

4.1.3 Regression analysis

The three reading time metrics (one from Natural Stories and two from OneStop) were modeled separately. The dependent measure was the by-token reading time, obtained by taking the mean residual reading times from a linear mixed model $\log(RT) \sim 1 + (1 \mid \textit{participant})$.

[5]The GPT-2 surprisal values were calculated using the code provided by (Oh and Schuler, 2023).

In each analysis, linear regression models were fitted to the reading times and evaluated by 10-fold cross-validation repeated 50 times. The predictive power of the models was quantified by their per-word mean squared error (MSE) of their prediction for the unseen data. The predictive power of a particular variable of interest was evaluated by comparing the MSEs of two models with and without that variable using the permutation test with $\alpha = 0.05$. Since we test each layer separately, p-values were adjusted for multiple comparisons by the Holm method. Linear models predicting RT_t at word w_t included variables associated not only with w_t but also with w_{t-2} and w_{t-1} to cover spillover effects, i.e., delayed reading slowdown effects due to the cost of preceding words, observed in many reading studies (Wilcox et al., 2023).

We considered four kinds of regression models, which ask different questions:

- Without any control variables: Does $\bar{\Delta}_t$ align with human reading times?
- With low-level (i.e., non-surprisal) variables: Does $\bar{\Delta}_t$ explain variance not explained by low-level features of the linguistic stimuli?
- With low-level variables and Mamba surprisal: Does $\bar{\Delta}_t$ explain variance even when predictability effects (Hale, 2001; Levy, 2008) are taken into account?
- With low-level variables and GPT-2 surprisal: Does $\bar{\Delta}_t$ explain variance even when the state-of-the-art predictor of predictability effects (Oh and Schuler, 2023) is taken into account?

We also asked whether these layers predict reading times independently. We searched for the combination of layers with the best predictive performance. Starting from the layer with the highest ΔMSE in the layer-wise evaluation, we considered one model at a time for the inclusion to the regression model, and included only when its inclusion resulted in a significant improvement of ΔMSE.

4.2 Results

The results with Natural Stories and Mamba-130M are shown in detail in Table 1. The results across datasets and models are summarized in Table 2. The general tendencies are similar across datasets and models. We find that $\bar{\Delta}_t$ from most layers are predictive of human reading times when it is the sole predictor in the regression model.

Baseline→	Intercept-only			Low-level variables			Low-level + Mamba surprisal			Low-level + GPT-2 surprisal		
↓Layer	ΔMSE	R^2	β	ΔMSE	R^2	β	ΔMSE	R^2	β	ΔMSE	R^2	β
0	0.01	0.00		0.05^{***}	0.49	$++-$	0.07^{***}	0.53	$++-$	0.06^{***}	0.54	$++-$
1	0.06^{**}	0.01	$+++$	0.02^{*}	0.49	$++-$	0.03^{***}	0.53	$++-$	0.03^{***}	0.53	$++-$
2	0.05^{**}	0.01	$+++$	0.02^{*}	0.49	$++-$	0.03^{***}	0.53	$++-$	0.03^{**}	0.53	$++-$
3	0.00	0.00		0.02^{*}	0.49	$++-$	0.01	0.53		0.01	0.53	
4	0.00	0.00		0.00	0.49		0.00	0.52		0.00	0.53	
5	0.02	0.00		0.01	0.49		0.01	0.53		0.01	0.53	
6	0.07^{***}	0.01	$++-$	0.01	0.49		0.00	0.52		0.00	0.53	
7	0.16^{***}	0.02	$+++$	0.01	0.49		0.00	0.52		0.00	0.53	
8	0.03^{*}	0.00	$-+-$	0.02^{**}	0.49	$-+-$	0.02^{*}	0.53	$-+-$	0.02^{*}	0.53	$-+-$
9	0.03^{*}	0.00	$-+-$	0.02^{*}	0.49	$-+-$	0.02^{*}	0.53	$-+-$	0.02^{*}	0.53	$-+-$
10	0.02^{*}	0.00	$-+-$	0.01	0.49		0.01	0.53		0.01	0.53	
11	0.04^{**}	0.01	$++-$	0.01	0.49		0.01	0.53		0.01	0.53	
12	0.11^{***}	0.02	$++-$	0.01	0.49		0.01	0.53		0.01	0.53	
13	0.04^{**}	0.01	$++-$	0.01	0.49		0.01	0.53		0.01	0.53	
14	0.07^{***}	0.01	$++-$	0.02^{*}	0.49	$-++$	0.00	0.52		0.00	0.53	
15	0.04^{**}	0.01	$++-$	0.00	0.49		0.00	0.52		0.00	0.53	
16	1.03^{***}	0.14	$+++$	0.06^{***}	0.49	$+++$	0.04^{***}	0.53	$-++$	0.04^{***}	0.53	$-++$
17	1.33^{***}	0.18	$+++$	0.09^{***}	0.50	$+++$	0.07^{***}	0.53	$-++$	0.07^{***}	0.54	$-++$
18	0.04^{**}	0.01	$++-$	0.01^{*}	0.49	$-++$	0.00	0.52		0.00	0.53	
19	0.06^{**}	0.01	$++-$	0.00	0.49		0.00	0.52		0.00	0.53	
20	0.16^{***}	0.02	$++-$	0.00	0.49		0.00	0.52		0.00	0.53	
21	0.22^{***}	0.03	$+++$	0.01	0.49		0.01	0.52		0.00	0.53	
22	0.58^{***}	0.08	$+++$	0.01	0.49		0.04^{***}	0.53	$+++$	0.04^{***}	0.53	$+++$
23	0.29^{***}	0.04	$+++$	0.00	0.49		0.00	0.52		0.00	0.53	

Table 1: Results of reading time modeling with Natural Stories and Mamba-130M. ΔMSE indicates the performance of $\bar{\Delta}_t$. The three signs in the β rows indicate the direction of the effect of $\bar{\Delta}_t$ on the ith word due to the Δ_t value of the ith word, the $(i-1)$th word, and the $(i-2)$th word, respectively. ΔMSE is calculated using cross validation, while R^2 and β are calculated by fitting a single regression model to the entire data. Significance code: $^{***}: p < 0.001, ^{**}: p < 0.01; ^{*}: p < 0.05$

The correlations between $\bar{\Delta}_t$ and human reading times are not generally high, but reach as high as $R^2 = 0.22$ (layer 41 of Mamba-2.8B, Natural Stories; cf. $R^2 = 0.21$ for GPT-2 surprisal in the same dataset). Across datasets and models, predictive power of some layers diminishes when control variables are taken into account, but some other layers remain significant. Also across datasets and models, multiple layers independently contribute to the prediction. Overall, these results indicate that $\bar{\Delta}_t$ is predictive of human reading times, and it explains a unique variance that is not explained by major predictors considered in the literature.

4.3 Follow-up analyses

We conducted post-hoc analyses to understand what makes the discretization timestep predictive of human reading times. Below, we focus on reading times from Natural Stories Corpus and Δ_t from Mamba-130M.

4.3.1 Layers' sensitivity to linguistic features

Correlations between linguistic features of w_t in the dataset and the discretization timestep $\bar{\Delta}_t$ are shown in Figure 3. Linguistic features include those used in the reading time regression, in addition to boolean flags indicating whether the word is at the beginning or at the end of the sentence, and the distance between the previous and the current words on the Penn TreeBank–style syntactic tree (Figure 4).

Two consistent trends are observed across layers: negative correlations between $\bar{\Delta}_t$ and the position of the word in the sentence, and positive correlations between $\bar{\Delta}_t$ and the beginning of the sentence. These correlations suggest that many layers undergo relatively large changes in their state at the beginning of the sentence, while they tend to maintain information within a sentence. Most layers also showed positive correlations between $\bar{\Delta}_t$ and syntactic distance. That is, $\bar{\Delta}_t$ of these layers increase when a large syntactic constituent completes and/or a new one begins (see Figure 4). These correlations suggest that the timestep sizes in the evolution of states in Mamba are sensitive to basic linguistic features.

There are also distinct patterns found in particular layers, suggesting some division of labor be-

Model	Baseline→ ↓Data		Intercept-only $\#_{sig}$	Intercept-only $\#_{ind}$	Intercept-only R^2_{Max}	Low-level variables $\#_{sig}$	Low-level variables $\#_{ind}$	Low-level variables R^2_{Max}	Low-level + Mamba surprisal $\#_{sig}$	Low-level + Mamba surprisal $\#_{ind}$	Low-level + Mamba surprisal R^2_{Max}	Low-level + GPT-2 surprisal $\#_{sig}$	Low-level + GPT-2 surprisal $\#_{ind}$	Low-level + GPT-2 surprisal R^2_{Max}
130M (24 layers)	NS	SPR	20	16	0.18_{17}	10	5	0.50_{17}	8	6	0.53_{0}	8	6	0.54_{17}
	OS	FP	24	19	0.04_{7}	23	11	0.19_{7}	24	11	0.20_{13}	21	5	0.21_{13}
		RP	20	16	0.01_{23}	11	5	0.08_{21}	16	7	0.09_{21}	20	5	0.10_{11}
2.8B (64 layers)	NS	SPR	64	25	0.22_{41}	29	21	0.51_{41}	32	22	0.53_{41}	26	20	0.54_{41}
	OS	FP	64	33	0.03_{2}	64	19	0.19_{18}	56	14	0.20_{40}	61	13	0.21_{42}
		RP	46	27	0.01_{60}	28	13	0.09_{47}	38	13	0.09_{47}	42	13	0.10_{47}

Table 2: Results of reading time modeling across datasets and model sizes. $\#_{sig}$ indicates the number of layers whose predictive power is significant. $\#_{ind}$ indicates the number of layers that are found to be independently predictive in the best model search. R^2_{Max} is the value of the regression model with the best-performing layer, with the subscript indicating the index of the best-performing layer. NS = Natural Stories; OS = OneStop; SPR = self-paced reading; FP = first pass time; RP = regression path duration.

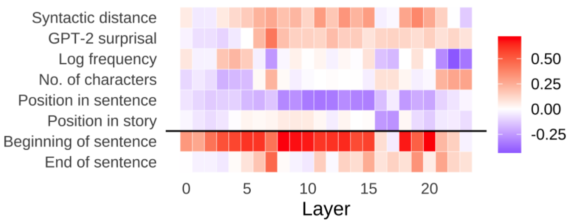

Figure 3: Correlations between linguistic features of w_t and the discretization timestep $\bar{\Delta}_t$. Correlations for the beginning and end of sentence flags are calculated using all words in the corpus; others are calculated only using words that are not at the beginning or at the end of the sentence.

tween layers. Layers 16 and 17 clearly differ from other layers in the linguistic features they are sensitive to. $\bar{\Delta}_t$ from these layers negatively correlate with the position of the word in the story, while their correlation with the position of the word in the sentence, the beginning of the sentence, and the syntactic distances are weaker than the surrounding layers. This means that these layers are not very sensitive to local information within a sentence, but rather maintain story-level information. The particularly high predictive power that these layers showed in the reading time modeling (see Table 1) suggests that there may be some parallel between the way these layers maintain story-level information and the way humans do.

4.3.2 Analysis of Long-range Dependencies through Interventions

As shown in Table 1, layers 16 and 17 exhibit particularly high predictive power for Natural Stories reading time. These layers show distinct sensitivity to linguistic features, suggesting a role in maintaining story-level information (Figure 3). To examine whether layers 16 and 17 actually capture long-range dependencies, we evaluate the contribution

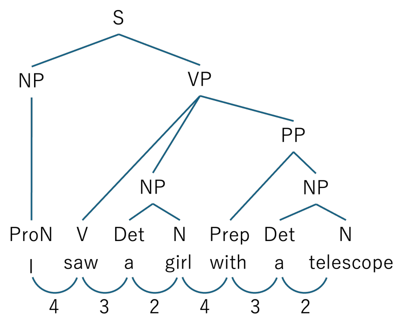

Figure 4: An example syntactic tree and the syntactic distance between adjacent words.

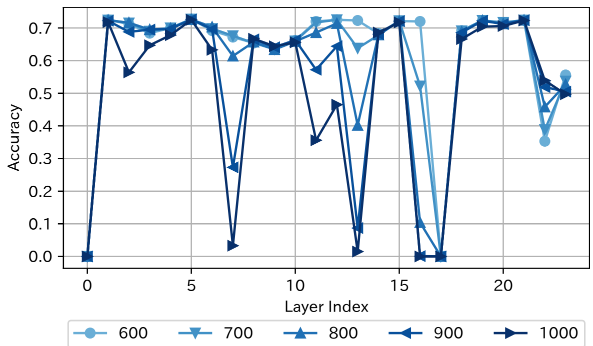

Figure 5: Accuracy in the passkey retrieval task when individually knocking out the SSM of each layer. The legend indicates the maximum token length of the input.

of each layer on the passkey retrieval task (Mohtashami and Jaggi, 2023) using an intervention-based analysis. The passkey retrieval task is commonly used to evaluate a model's ability to handle long-range dependencies. It assesses whether a model can correctly recover a specific string from an input consisting of that string embedded within a sequence of noise tokens.

Settings The experimental setup is as follows. The model is given an input that begins with a six-digit number, followed by a sequence of noise sentences, and finally a prompt instructing it to reproduce the passkey. Concretely, the input takes the following form: "The passkey is 317451. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again. ... The passkey is". The output is considered correct if it exactly matches the passkey provided at the beginning of the input. The distance between the passkey and the output position is controlled by repeating the noise sentences until a specified length is reached. In this experiment, the length of the noise was adjusted so that the total input length was approximately {600, 700, 800, 900, 1000} tokens. For each length setting, 1,000 input instances are generated, and accuracy is evaluated accordingly. We measured the performance of the Mamba models with one layer knocked out at a time, by setting the coefficient $\overline{B}_t$ for the input in Equation (1) to zero at all times, thereby ensuring that $\boldsymbol{h}_t = 0$ is always maintained.

Results The results are shown in Figure 5. As illustrated, a notable drop in accuracy is observed when layer 0, 7, 13, 16, or 17 is knocked out. In particular, even when the input length is reduced to 800 tokens, layer 16 maintains an accuracy in the 10% range, and for layers 0 and 17, the accuracy remains at 0% even when the input is shortened to 600 tokens.

Discussions The results suggest some layers are essential for inference that requires the transmission of information over long-range context. However, it should be noted that these layers are not necessarily specialized for long-range dependencies; it remains possible that they are simply essential for text processing in general. Interestingly, the layers 16 and 17 of Mamba-130M are not particularly predictive of OneStop reading time, unlike in the case of Natural Stories. One possible interpretation of this discrepancy is that human readers read differently depending on the reading paradigm and/or materials. For example, the self-paced reading paradigm might encourage readers to spend more time when the input word is associated with a distant context since the retrieval of context relies solely on memory, while in the eyetracking paradigm the retrieval can be done by physically looking at the context. This is a pure speculation, and we leave it for future work exactly why different layers are responsible for predicting reading times from different datasets.

5 Mamba as a New Lens on Human Language Processing

Having established an empirical connection between the discretization timestep $\bar{\Delta}_t$ and human reading times, we turn to theoretical discussions on how Mamba processes language over time with ever-updated memory. We suggest that the model can be informative to the science of language processing by humans, which also process language in an incremental fashion under memory constraints.

5.1 Transition Matrix Eigenspectrum as Memory Retention Ability

The extent to which Mamba attenuates memory of past context depends on the elements of the transition matrix $\exp(A)$. In this section, we first explain this property analytically and then verify that layers 16 and 17, which were effective for predicting reading time, also exhibit distinctive behavior in terms of the eigenvalues of the transition matrix.

Since the transition matrix $\exp(A)$ of Mamba is diagonal, its eigenvalues are simply its diagonal elements. In Equation (1), when the elements of the transition matrix are close to one, that is, when $\exp(A) \approx I$, we have $\overline{A}_t \approx I$ regardless of the value of Δ_t, and thus the past context $\boldsymbol{h}_{t-1}$ is retained over the long term. In contrast, when the elements of the transition matrix are close to zero, that is, when $\exp(A) \approx O$, we have $\overline{A}_t \approx O$, and the past context $\boldsymbol{h}_{t-1}$ is rapidly forgotten. In this way, the magnitude of the eigenvalues of the transition matrix $\exp(A)$ serves as an indicator for evaluating the ability of each SSM layer to retain information over long timescales, from a perspective distinct from the discretization timestep Δ_t.

Figure 6 shows the eigenvalue distributions for the entire layer and for layers 16 and 17, where timestep Δ_t does not react to the sentence boundary. From the figure, it can be seen that while the overall eigenvalue distribution peaks near zero, the eigenvalue distributions of layers 16 and 17 show a higher proportion of eigenvalues near 0.5 compared to the overall distribution. The eigenvalue analysis of layers 16 and 17 from the perspective of SSM dynamics thus agrees with the analysis based on linguistic features and Δ_t in suggesting that these layers tend to retain information over the long term.

In short, layers of Mamba process information at different timescales via different eigenvalues of transition matrices (as well as varying Δ_t), which can be interpreted as *memory retention ability* of

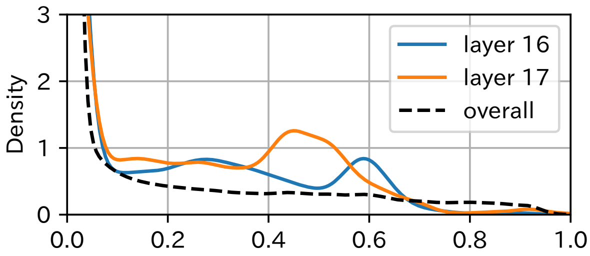

Figure 6: Eigenvalue distribution of the transition matrix $\exp(A)$. This plots the eigenvalue distribution over all layers and separately for layers 16 and 17, where timestep Δ_t does not react to sentence boundaries.

each layer. Interesting future directions based on this observation include comparing this mechanism to human brain activities, which also seem to keep track of information that unfolds at different timescales (Gwilliams et al., 2024), with potential division of labor between neural populations (Gwilliams et al., 2022).

5.2 Discretization Timestep as State Transition Uncertainty

We theoretically show that the component Δ_t corresponds, in a sense, to the *uncertainty* in state transitions.

We now present the following formal statement. First, to interpret the hidden-state transitions of Mamba as a probabilistic model, we formally introduce a noise term. Specifically, while the original transition function is given by Equation (1), we add a small Gaussian noise term $w(t)$ to it as follows:

$$h'(t) = A(t)h(t) + B(t)x(t) + w(t) \quad (6)$$

$$w(t) \sim \mathcal{N}(0, Q_c). \quad (7)$$

This allows the hidden-state transition $p(h_t \mid h_{t-1})$ to be written in closed form:

$$p(h_t \mid h_{t-1}) := \mathcal{N}\left(h_t\,;\,\overline{A}_t h_{t-1} + \overline{B}_t x_t, Q_d\right), \quad (8)$$

where Q_d denotes the noise intensity after discretization. Under this standard assumption, it follows that *the conditional entropy (uncertainty) of the hidden-state transition $\boldsymbol{h}_t$ is determined by Δ_t, with the uncertainty increasing as Δ_t becomes larger*:

$$H[h_{i+1} \mid h_i] = \frac{1}{2}\sum_{i=1}^{n} \log\left(e^{2A_{ii}\Delta_t} - 1\right) + \text{const.}, \quad (9)$$

where A_{ii} denotes the i-th diagonal element of the transition matrix A before discretization. For a detailed derivation, refer to Appendix §A.

In practice, how does Mamba adjust the timestep Δ_t in response to the given input text? Figure 2 shows that in several layers, the timestep Δ_t peaks at the beginning of sentence tokens (this trend is also quantitatively confirmed in the analysis in §4.3.1 (Figure 3)). Mamba increases timestep Δ_t at the beginning of sentences, precisely at the points where it is difficult to predict from the immediate preceding context (typically ending with periods). As confirmed in this section, it is intuitive that a larger timestep Δ_t corresponds to greater uncertainty in state transition. However, it is not trivial that Mamba possesses the property of incrementing Δ_t at uncertain points during state transition. Why and how this property was acquired through learning remains an open question. Finally, it should be noted that the considerations here are based on the interpretation when a noise term is introduced into Mamba and do not naturally follow from the original deterministic model.

It should be noted that entropy/uncertainty discussed here is *not* directly related to the uncertainty about subsequent input much discussed and investigated in the cognitive modeling literature (Hale, 2016). Instead, we are looking at uncertainty about the state that model arrive at when the processing of the current word is complete. This notion of uncertainty may offer a new perspective on the influence of noise on language processing, a hotly discussed topic in computational psycholinguistics. In existing models of noisy language processing (Futrell et al., 2020; Hahn et al., 2022), the noise erases individual words at a certain probability when each word is read. In the current assumption, noise is added to the spatial representation of the context and accumulates over time. Future work can explore whether such a view on noise better capture limitations and rationality of human language processing.

6 Related Works

Kuribayashi et al. (2025) performs cognitive modeling using internal components of language models, although it focuses on Transformer-based models rather than SSM-based ones. The authors demonstrate an intriguing correspondence between the ordering of layers and the temporal order of human responses: earlier layers align with early eyemovement measures, while later layers correspond to later responses associated with deeper semantic processing. Their approach converts intermediate layer representations into values corresponding to surprisal and uses them for cognitive modeling. In contrast, our study directly incorporates the internal model component Δ_t, interpretable as the processing time of an SSM-based language model, into reading-time modeling, which constitutes a key difference from their approach.

Sharma et al. (2024) and Endy et al. (2025) propose methodologies for identifying layers that are responsible for transmitting factual information. Specifically, given the input "Michael Jordan professionally played," they examine how the factual knowledge "basketball" is propagated through the model to the final output token. Through intervention experiments that disrupt information transmission pathways, both studies demonstrate that factual information is conveyed from the subject to the final token primarily via later intermediate layers. These approaches appear promising as follow-up analyses for reading-time modeling, as they offer ways to probe what computations are performed in individual layers. However, their analyses focus on short factual sentences and do not address the processing of story-level long-form text as considered in this work; extending such analyses to longer narratives remains an open problem for the field.

Why, in the first place, do the discretization timesteps Δ_t peak at sentence boundaries? Liu and Ding (2025) provides insights that help address this question. They report that large language models, when processing information, constrain the scope of information integration at sentence boundaries, similarly to humans. Their experiments focus on Qwen and Pythia models with attention mechanisms, which can attend to the entire input text. Nevertheless, they show that these models tend to preferentially rely on words within the same sentence or clause. The phenomenon observed in Figures 2 and 3 of our study, where peaks of timesteps Δ_t arise at sentence boundaries, likewise induces an effect that limits the range of information integration at such boundaries, consistent with their findings. This is because timestep Δ_t functions as a gate: when it becomes large, it forgets past memory h_{t-1} and incorporates more of the current input x_t. However, layer 17 of Mamba-130M, which showed the highest predictive performance for Natural Stories in the current experiment, does not exhibit peaks at sentence boundaries. This suggests that information transmission across sentences is an essential component explaining reading behavior, and thus our result differs in this respect from theirs.

7 Conclusion

We showed that Mamba's discretization timestep Δ_t is predictive of human per-word reading times. The predictive power of discretization timestep Δ_t from the best-performing layer is comparable to that of GPT-2 surprisal, and Δ_t from many layers remain significant even when GPT-2 surprisal and other known predictors are controlled for. We also conducted formal analyses of Mamba's architecture and internal dynamics, and suggested that Mamba can serve as a new, valuable lens to look at human real-time language processing with ever-updated memory. With this new lens, future work may investigate how humans, who are constrained by the flow of time and ever-changing memory, still show remarkable ability to use language that unfolds over time, often quite rapidly.

8 Limitations

We have not clarified how the model learns to increase discretization timestep Δ_t at sentence boundaries or when the syntactic path length to adjacent words is long. Therefore, we report only that we discovered and examined that the timestep Δ_t indeed contributes to the reading time modeling. To understand how the timestep Δ_t peaks at sentence boundaries and other locations, statistical analysis of the training corpus and analysis of the learning dynamics would be necessary, but this is beyond the scope of this paper.

The analysis in Section §4.3.2 provided evidence that layer 17 of Mamba-130M, which showed the strongest contribution to predicting Natural Stories reading time, potentially captures long-range dependencies. However, this does not directly imply that long-range dependencies are inherently a crucial factor for modeling reading time. In other words, a principled explanation for why the timestep Δ_t contributes to reading time modeling has yet to be established.

Since experiments were conducted solely using Mamba, the discussions in this paper cannot be applied to other architectures. The reason only one architecture was adopted is that Mamba is the only one where the model itself dynamically determines the discretization timestep based on SSM.

Acknowledgments

This work was supported by JST SPRING (JPMJSP2104), JST FOREST (JPMJFR2331), and JSPS KAKENHI (JP22H05106).

This study was inspired by discussions held at the weekly "317 Café" at the National Institute for Japanese Language and Linguistics (NINJAL).[6] We would like to express our sincere gratitude to all participants of the 317 Café.

References

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. Onestop: A 360-participant English eye tracking dataset with different reading regimes. *Scientific Data*, 12(1995).

Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, and 1 others. 2025. Nvidia nemotron 3: Efficient and open intelligence. *arXiv preprint arXiv:2512.20856*.

Guy E Blelloch. 1990. Prefix sums and their applications.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:192–210.

Nir Endy, Idan Daniel Grosbard, Yuval Ran-Milo, Yonatan Slutzky, Itay Tshuva, and Raja Giryes. 2025. Mamba knockout for unraveling factual information flow. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23457–23477, Vienna, Austria. Association for Computational Linguistics.

Richard Futrell, Edward Gibson, and Roger Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. 2022. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, 13(1):6606.

[6] https://sites.google.com/view/317cafe/

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Remi King. 2024. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *bioRxiv*.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *PNAS*, 119(43).

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10:397–412.

Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228–238.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, and 1 others. 2025. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Wei Liu and Nai Ding. 2025. Information integration in large language models is gated by linguistic structural markers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6903–6915, Suzhou, China. Association for Computational Linguistics.

David Marr. 1982. *Vision*. WH Freeman, San Fransisco, CA.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 54567–54585. Curran Associates, Inc.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Preferred Networks, Kaizaburo Chubachi, Yasuhiro Fujita, Shinichi Hemmi, Yuta Hirokawa, Kentaro Imajo, Toshiki Kataoka, Goro Kobayashi, Kenichi Maehashi, Calvin Metzger, and 1 others. 2025. Plamo 2 technical report. *arXiv preprint arXiv:2509.04897*.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Arnab Sen Sharma, David Atkinson, and David Bau. 2024. Locating and editing factual associations in mamba. In *First Conference on Language Modeling*.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Songlin Yang and Yu Zhang. 2024. Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism.

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2025. If attention serves as a cognitive model of human memory retrieval, what is the plausible memory representation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9795–9812, Vienna, Austria. Association for Computational Linguistics.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. 2024. Falcon mamba: The first competitive attention-free 7b language model. *arXiv preprint arXiv:2410.05355*.

A Derivation of State Transition Uncertainty

In this section, we derive Equation (9). The state equations for a continuous-time SSM with a Gaussian noise term $w(t)$ are as follows:

$$h'(t) = A(t)h(t) + B(t)x(t) + w(t) \tag{10}$$
$$w(t) \sim \mathcal{N}(0, Q_c), \tag{11}$$

where Q_c denotes the noise intensity in continuous time and is assumed to be diagonal, similar to the transition matrix A of Mamba. This is equivalent to interpreting the state-space model as the following probabilistic model:

$$p(h_t \mid h_{t-1}) := \mathcal{N}\left(h_t \,;\, \overline{A}_t h_{t-1} + \overline{B}_t x_t, Q_d\right), \tag{12}$$

where $\overline{A}_t$ and $\overline{B}_t$ denote the SSM parameters after discretization (Eq. (3) and (4)), and Q_d denotes the noise intensity after discretization.

Applying Zero-Order Hold (ZOH) discretization yields the following noise term:

$$w_t \sim \mathcal{N}(0, Q_d), \;\; Q_d = \int_0^{\Delta_t} e^{A\tau} Q_c e^{A^\top \tau} d\tau, \tag{13}$$

where Q_d is the discrete representation of the continuous-time noise intensity Q_c at the sampling interval Δ_t, and is a diagonal matrix. Since the transition matrix e^A of Mamba is diagonal, the integrand is also a diagonal matrix, allowing Equation (13) to be solved as follows:

$$(Q_d)_{ii} = \int_0^{\Delta_t} e^{2A_{ii}\tau}(Q_c)_{ii} d\tau = (Q_c)_{ii} \frac{e^{2A_{ii}\Delta_t} - 1}{2A_{ii}}. \tag{14}$$

The uncertainty before and after the state transition can be quantified by the conditional differential entropy $H[h_t \mid h_{t-1}]$:

$$H[h_t \mid h_{t-1}] = -\int_{\mathbb{R}^n} p(h_t \mid h_{t-1}) \log p(h_t \mid h_{t-1}) dh_t. \tag{15}$$

Here, by performing the following change of variables, $\mathcal{N}(h_t \,;\, \overline{A}_t h_{t-1} + \overline{B}_t x_t, Q_d)$ is equivalent to $\mathcal{N}(w_t \,;\, 0, Q_d)$:

$$\mathcal{N}(h_t \,;\, \overline{A}_t h_{t-1} + \overline{B}_t x_t, Q_d) \tag{16}$$
$$= \frac{1}{\sqrt{(2\pi)^n |Q_d|}} \exp\left(-\frac{1}{2}(h_t - (\overline{A}_t h_{t-1} + \overline{B}_t x_t))^\top Q_d^{-1} (h_t - (\overline{A}_t h_{t-1} + \overline{B}_t x_t))\right) \tag{17}$$
$$= \frac{1}{\sqrt{(2\pi)^n |Q_d|}} \exp\left(-\frac{1}{2} w_t^\top Q_d^{-1} w_t\right) \tag{18}$$
$$= \mathcal{N}(w_t \,;\, 0, Q_d). \tag{19}$$

Then, Equation (15) can be rewritten as follows from the above equation and Equation (12):

$$H[h_t \mid h_{t-1}] = -\int_{\mathbb{R}^n} \mathcal{N}(w_t \,;\, 0, Q_d) \log \mathcal{N}(w_t \,;\, 0, Q_d) dw_t. \tag{20}$$

Since the right-hand side is the differential entropy of a Gaussian distribution, it is given by the following:

$$H[h_t \mid h_{t-1}] = \frac{1}{2} \log\left((2\pi e)^n |Q_d|\right) \tag{21}$$

Here, substituting the components of Q_d derived earlier in Equation (14), we obtain the following expression:

$$H[h_{i+1} \mid h_i] = \frac{1}{2} \sum_{i=1}^{n} \log \left(e^{2A_{ii}\Delta_t} - 1 \right) + \text{const}. \tag{22}$$

where the terms independent of timestep Δ_t are grouped as a constant term.

Therefore, the uncertainty in state transitions monotonically increases with respect to the discretization timestep Δ_t. Conversely, as the timestep Δ_t approaches zero, the differential entropy tends to $-\infty$, which indicates the absence of uncertainty. This leads to the intuitive conclusion: if the sampling period is coarse, the uncertainty regarding the next state increases; conversely, if sampled finely, the uncertainty becomes sufficiently small.