Spring 5-10-2024

Temporal Order Memory in Naturalistic Events Is Influenced by Semantic Knowledge and Hierarchical Event Structure

Yining Ding
Washington University in St. Louis

Recommended Citation

Ding, Yining, "Temporal Order Memory in Naturalistic Events Is Influenced by Semantic Knowledge and Hierarchical Event Structure" (2024). *Arts & Sciences Electronic Theses and Dissertations*. 3096.
https://openscholarship.wustl.edu/art_sci_etds/3096

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological & Brain Sciences

Temporal Order Memory in Naturalistic Events Is Influenced by Semantic Knowledge and Hierarchical Event Structure

by

Yining Ding

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my thesis committee, Dr. Jeffrey Zacks, Dr. Ian Dobbins, and Dr. Zachariah Reagh, for their insights and guidance on this project. A special thanks goes to my advisor Dr. Jeffrey Zacks for his amazing mentorship and support during my journey in graduate school.

I am grateful to DCL and CML lab members for helpful discussions and the cherished time we spent together both inside and outside the lab. I would like to thank Devon Alperin for her help on stimuli development and recall coding in this project.

Finally, I would like to express my gratitude to my parents, family, and friends for their love and support along the way.

Yining Ding

Washington University in St. Louis

May 2024

ABSTRACT OF THE THESIS

Temporal Order Memory in Naturalistic Events Is Influenced by Semantic Knowledge and Hierarchical Event Structure

by

Yining Ding

Master of Arts in Psychological & Brain Sciences

Washington University in St. Louis, 2024

Professor Jeffrey. M. Zacks, Chair

As people go through everyday life, they segment their continuous sensory experience into distinct events, and the event structure being perceived during encoding has important implication on how event dynamics is later reconstructed from long-term memory. Previous studies using discrete pictorial stimuli showed that people are sometimes better at remembering the temporal order of items occurring within the same perceptual context than items spanning across a perceptual boundary, but that the opposite can also occur if contextual cues are available during retrieval. However, given that these paradigms only tested the episodic memory of arbitrary temporal associations, it is unclear if the conclusions can be generalized to everyday scenarios that rely heavily on structured event knowledge for perception and memory. In the current study, we developed a set of hierarchically organized narrative stimuli describing everyday events, with semantic order constraints among events either on the coarse-level or on the fine-level. In Experiment 1 and 2, we found that within-event temporal order memory was improved when fine-level semantic constraints were provided, and across-event temporal order memory was improved when coarse-level semantic constraints were provided. We observed these effects after both a shorter (2.5-minute) and a longer (20-minute) delay, which demonstrated that participants could use semantic order constraints on either coarse- or fine-level to facilitate temporal order reconstruction. In Experiment 3, we tested serial recall of the narratives and found that participants frequently chucked their recall based on coarse-level event membership. Together, these results suggested that temporal order memory of everyday events should be primarily viewed as a reconstruction process that utilizes multiple sources of information apart from episodic memory.

Chapter 1: Introduction

Human experience unfolds unidirectionally over time, yet people's ability of remembering past experiences resembles a form of "mental time travel" that resist the irreversibility of time (Tulving, 2002). A successful reconstruction of the past requires remembering not only "what" happened, but also "when" it happened.

The representations people form when perceiving these experiences have important implication on how we later remember their content and reconstruct the relationship among them. Event Segmentation Theory (EST) suggests that during the ongoing perception of everyday activities, people spontaneously segment their continuous sensory experience into distinct events (Zacks et al., 2007). According to EST, people maintain an active event model in working memory that describes the current situation. The current event model is constructed based on a combination of sensory information available in the immediate environment and event schemas that represent existing semantic knowledge about the event. People use the stable event model to make predictions about what is going to happen next, and adaptively update their event model when there is a transient increase in prediction errors. When event model update occurs, people typically segment the event and perceive an event boundary. Supporting this theoretical framework, there is substantial evidence suggesting that people agree on when event boundaries occur and are sensitive to hierarchical structures in the events (Zacks et al., 2001; Zacks et al., 2006; Sasmita & Swallow, 2022). The perception of "partonomic hierarchy" in events, which is defined as "fine-grained events clustering into larger coarse-grained events" (Radvansky & Zacks, 2014), has been shown to increase as participants became more familiar with an event sequence (Hard et al., 2006; Zacks, 2020). In both reading and film viewing, it has been demonstrated that people are able to constantly track multiple dimensions in the story, including characters, spatial locations, goals, and the causal and temporal relations, to segment and update their current event model when incoming information no longer matches predictions generated by the current model (Gernsbacher, 1991; Zwaan & Radvansky, 1998; Zacks et al., 2009). For example, when a reader encounters a sentence in a story saying that the main character goes from their home to the school, the change in spatial location suggests that the reader needs to update their previous event model about home activities to accommodate new events that happen at the school. One important behavioral signature for event model updating at event boundaries is that reading time increases when spatial shifts (Radvansky et al., 2001; Zwaan & Radvansky, 1998), temporal shifts (Radvansky & Copeland, 2010; Rinck & Weber, 2003), goal shifts (Radvansky et al., 2001), character shifts (McNerney et al., 2011), or causal shifts (Radvansky et al., 2001; McNerney et al., 2011) occur in narratives.

Event structure influences how events are encoded, stored, and later reconstructed from long-term memory (Radvansky, 2012; Radvansky & Zacks, 2017; Zacks, 2020; Rubin & Umanath, 2015). In one study conducted by Ezzyat & Davachi (2011), they asked participants to read narratives that contain event boundary sentences indicating temporal shift in the storyline. When they later cued people with a sentence from the narrative and asked them to recall the next sentence, they found that the cued recall performance was worse if there was an event boundary separating the cue and the to-be-recalled sentence, and the effect was later replicated in both younger and older adults (Davis & Campbell, 2023). This evidence suggests that event boundaries help discretize experience into distinct episodes in long-term memory, which reduces interference across different event models during retrieval.

Consistent with the finding that boundaries seem to impair cued-recall performance, there have been studies showing that contextual boundary disrupts people’s ability to remember the temporal order among items. Unlike previous studies using narratives as stimuli, these paradigms aimed to simplify naturalistic event structure and demonstrate simple contextual changes are sufficient to cause event segmentation that supports better within-event associative memory (Heusser et al., 2018). In these paradigms, participants were presented with a series of discrete pictures during encoding, and the “context” in which these items appeared changed periodically to create perceptual boundaries. At the retrieval time, they were given two pictures and were asked to judge the relative recency of these two pictures (“Which of these two stimuli were seen first?”), as well as rate the subjective temporal distance between them during encoding (“How far apart in time were the two stimuli presented?”). A consistent finding is that people are more likely to forget the order of the two probed pictures, if they are encoded in two different contexts comparing to in a uniform context, when contextual change are operationalized by a change in stimuli category (DuBrow & Davachi, 2013, 2014; Sols et al., 2017), spatial location (Horner et al., 2016; Gurguryan et al., 2021), background color (Heusser et al., 2018; Pu et al., 2022), background sound (Clewett et al., 2020; McClay et al., 2022), encoding task (Wang & Egner, 2022), or the magnitude of reward prediction error (Rouhani et al., 2020). This boundary-related disruption in temporal order memory is often accompanied by an inflation in temporal distance judgment, which means that participants are more likely to rate stimuli spanned across event boundaries as temporally further away from each other, even though the actual temporal lag in between remains constant across conditions (DuBrow & Davachi, 2013; Ezzyat & Davachi, 2014).

Together, these findings can be interpreted by two different associative mechanisms for how temporal information is stored in memory: One is the chaining theory that emphasizes using direct item-item associations for recency judgment (Lewandowsky & Murdock, 1989; Murdock, 1983). According to this account, memory for the serial order of items is supported by encoding and retrieving the pairwise associations between sequential items (Lewandowsky & Murdock, 1989). This predicts that event boundaries induced by contextual changes will disrupt the formation of associative links between items during encoding, thereby causing recency information across events more difficult to retrieve (Heusser et al., 2018). Another associative account is given by the temporal context model (TCM) and the related context maintenance and retrieval (CMR) model, which emphasizes an indirect temporal linking mechanism among items through their shared, gradually changing temporal context (Howard & Kahana, 2002; Polyn et al., 2009). According to this account, event boundaries may cause an abrupt shift in the slowly drifting temporal context representation, which makes it easier to retrieve the temporal association among items within the same event than across different events (DuBrow & Davachi, 2013; DuBrow et al., 2017).

There is also empirical evidence suggesting that whether event boundaries exert a disruptive effect on temporal order memory depends on other factors. For example, Wen and Egner (2022) demonstrated that if the encoding context of items was salient enough and available during retrieval, recency judgment for items spanning across two events would be more accurate than items within the same event, and this effect co-occurred with an inflated temporal distance rating for across-event items. This finding contradicts the pattern predicted by frameworks centered on associative mechanisms, and instead provides partial support for distance-based mechanisms for memory of time (Hintzman, 2005). Distance-based theories suggested that recency judgment was operationalized through a comparison of the relative memory strength of each individual item. According to this account, recency judgment should be easier for items that are relatively farther apart in time, due to their more salient difference in strength.

To summarize, recent studies using picture-list learning paradigm yielded mixed results in identifying the mechanism underlying temporal order memory. However, one may argue that the mental process we typically rely on for remembering the order of everyday events operates somewhat differently from what we use for remembering the order of newly encountered, arbitrary pictures in a laboratory setting. As Friedman (1993) argued, our ability to remember the temporal relationship among autobiographical events should be primarily viewed as a reconstruction process, which specializes on combining episodic information with general knowledge about recurring time patterns to make inferences. This coincides with the viewpoint from Bartlett (1932) that structured knowledge about events, which are known as schemas, play an important role in how we reconstruct our memories for specific life episodes.

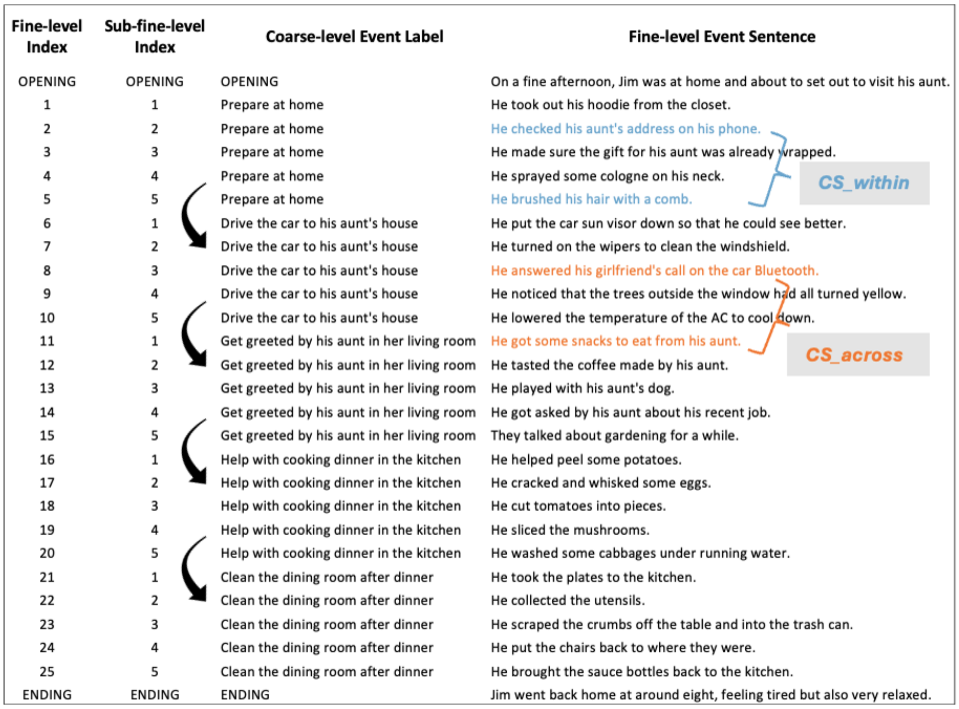
Indeed, multiple empirical studies have shown that people have semantic knowledge about stereotyped event sequences in their long-term memory, which are referred to as "scripts" (Abelson, 1981). When people describe what typically happens during familiar activities (e.g., eating in a restaurant), they have good agreement on the fine-level events that constitutes the coarse-level event (e.g., entering, ordering, eating, etc.), as well as specific characters (e.g., customer, waiter, cashier, etc.) and actions (e.g., pick up menu, look at menu, etc.) (Bower et al., 1979). However, scripts also differ in terms of the regularity of the temporal order among events. (Abelson, 1981; McRae et al., 2021). On one end of the spectrum, some scripts have strong constraints on the ordering of their constituent events. These constraints sometimes result from causal relations (e.g., Eating a meal can only happen after ordering the meal), and sometimes come from socio-cultural conventions (e.g., In western culture, the main course is typically served after the appetizer, and dessert is typically served after the main course). On the other end of the spectrum, some scripts have constituent activities that are generally agreed-upon, but with weak constraints on the way they are sequenced (e.g., An event like cleaning a room contains typical subevents like vacuuming the floor and clean the table, but they can happen in any order). Key aspects of semantic knowledge about familiar events are (a) what fine-level events are contained in a coarse-level event, and (b) the typical order events at a given grain. Both aspects of semantic knowledge may serve as important sources of information for how we encode and reconstruct the temporal dynamic in everyday life. This proposal would be challenging to test using picture-list learning paradigms because the order of both item and context changes in these paradigms are arbitrary, and because the relationship between items and contexts are frequently orthogonalized. In other words, most of the associations formed and tested in these paradigms are one-shot learned episodic associations, which prevents participants from using existing semantic knowledge to facilitate remembering.

The goal of the current study was to investigate the extent to which semantic knowledge about event structure facilitates the encoding and retrieval of temporal order relationships among events. We would like to reconceptualize the role event boundary plays in temporal order memory while considering how it interacts with existing schematic information to facilitate event model construction. To this end, we created narratives about everyday activities with a two-level hierarchical structure, which paralleled the stimuli design in well-established picture-list learning paradigms. This was illustrated in Figure 1. In our narratives, “Fine-level events” were similar to “items,” and “Coarse-level events” were similar to “contexts.” But instead of having random item-context pairing (e.g., a picture of a ball shown in a purple background) as a new episodic association to learn, the membership of fine-level events belonging to certain coarse-level events was designed to rely on participants' existing semantic knowledge (e.g., The fine-level event "*Peeling some potatoes*" belongs to the coarse-level event "*Help with cooking dinners in the kitchen*"). In addition, instead of having completely random ordering among "items" and "contexts," the narratives were constructed to have strong order constraints at one of the two levels: For Coarse-level Semantic (*CS*) narratives, there were strong semantic order constraints only among coarse-level events, but there were no semantic order constraints among each set of fine-level events within each coarse-level event. For example, in the "*visiting aunt*" CS narrative (see Figure 1(A), on the coarse-level, most people would agree that the protagonist would first prepare at home, then drive the car, and then get greeted by his aunt in her living room. But within the "*preparing at home*" coarse-level event, it would make sense for the protagonist to do things like taking out hoodie and checking address in any order). Fine-level Semantic (*FS*) narratives have a structure that was the complement of CS narratives: there were strong semantic order constraints on each set of fine-level events within each coarse-level event, but there were no semantic order constraints among coarse-level events. For example, in the "*visiting the zoo*" FS narrative (see Figure 1(B), on the coarse-level, it would make sense for the protagonist to do things like getting a tattoo or watching the sea lion show in any order. But within the "*visit the snack cart*" coarse-level event, most people would agree that the protagonist would first wait in the line, then tell the owner what he wants, and then pay and get food.

We hypothesized that when reading each narrative, changes in coarse-level events would be perceived as event boundaries and would induce typical boundary-related effects, including increased reading time for boundary sentences during encoding (Radvansky et al., 2001; Zwaan

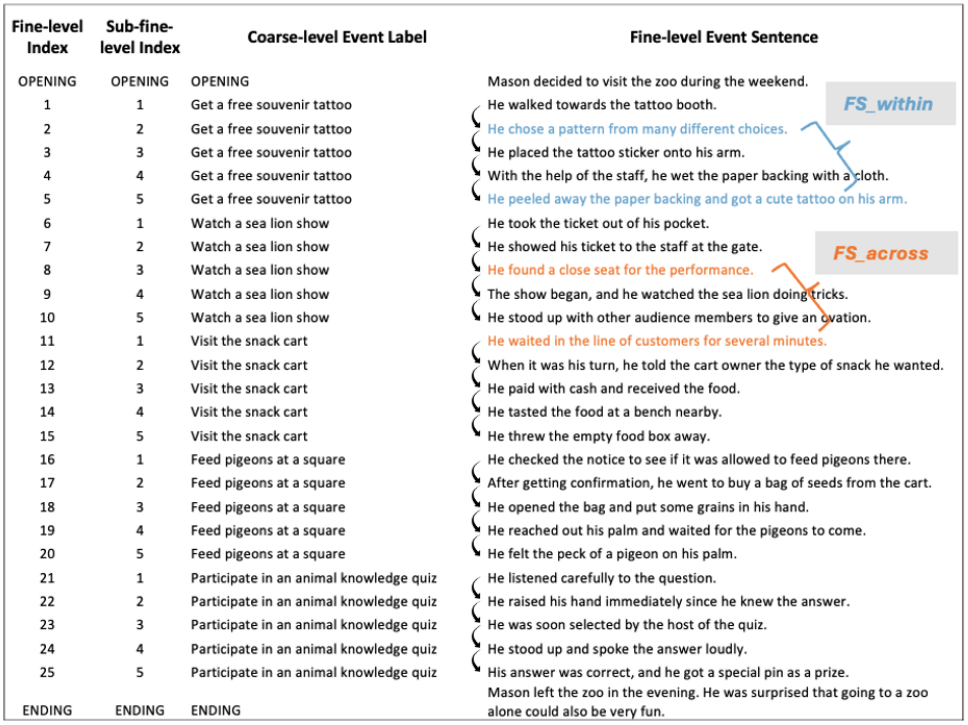
& Radvansky, 1998) and inflated temporal distance rating for fine-level event pairs spanning across event boundaries during retrieval (DuBrow & Davachi, 2013; Ezzyat & Davachi, 2014). Critically, we hypothesized that the relative accuracy of recency judgment should depend on where semantic order constraint is present: There should be better across-event temporal order memory when only coarse-level semantic constraints were provided (in CS narratives), but better within-event temporal order memory when only fine-level semantic constraints were provided (in FS narratives). These effects should be present both during a relatively short delay between encoding and retrieval (about 2.5 minutes, Experiment 1) and a longer delay (about 20 minutes, Experiment 2). In addition, we hypothesized that the temporal organization of recalled events should be influenced by semantic order knowledge and hierarchical event structure (Experiment 3).

A. Sample Coarse-level Semantic (CS) Narrative



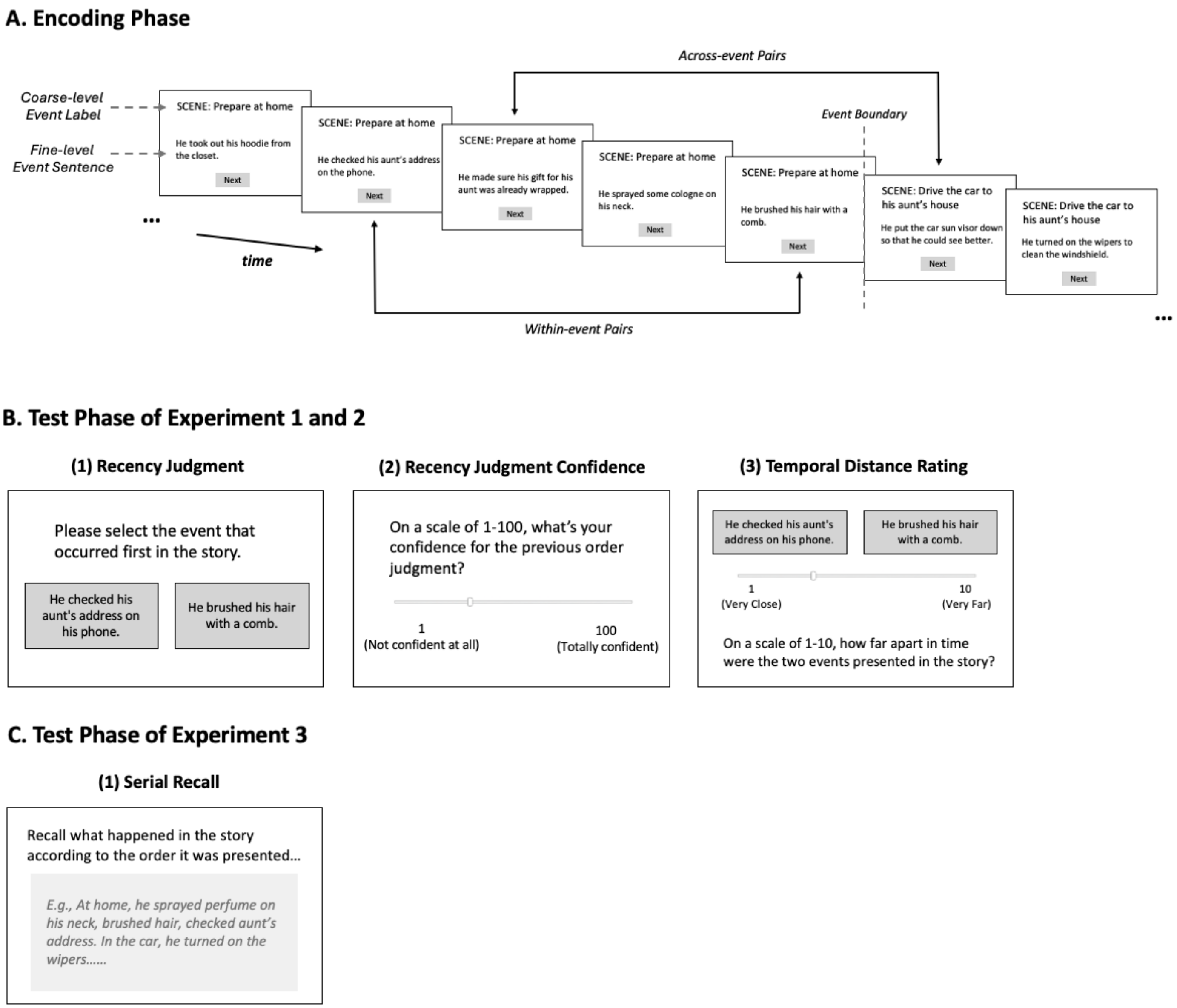
Fine-level Index	Sub-fine-level Index	Coarse-level Event Label	Fine-level Event Sentence
OPENING	OPENING	OPENING	On a fine afternoon, Jim was at home and about to set out to visit his aunt.
1	1	Prepare at home	He took out his hoodie from the closet.
2	2	Prepare at home	He checked his aunt's address on his phone.
3	3	Prepare at home	He made sure the gift for his aunt was already wrapped.
4	4	Prepare at home	He sprayed some cologne on his neck.
5	5	Prepare at home	He brushed his hair with a comb.
6	1	Drive the car to his aunt's house	He put the car sun visor down so that he could see better.
7	2	Drive the car to his aunt's house	He turned on the wipers to clean the windshield.
8	3	Drive the car to his aunt's house	He answered his girlfriend's call on the car Bluetooth.
9	4	Drive the car to his aunt's house	He noticed that the trees outside the window had all turned yellow.
10	5	Drive the car to his aunt's house	He lowered the temperature of the AC to cool down.
11	1	Get greeted by his aunt in her living room	He got some snacks to eat from his aunt.
12	2	Get greeted by his aunt in her living room	He tasted the coffee made by his aunt.
13	3	Get greeted by his aunt in her living room	He played with his aunt's dog.
14	4	Get greeted by his aunt in her living room	He got asked by his aunt about his recent job.
15	5	Get greeted by his aunt in her living room	They talked about gardening for a while.
16	1	Help with cooking dinner in the kitchen	He helped peel some potatoes.
17	2	Help with cooking dinner in the kitchen	He cracked and whisked some eggs.
18	3	Help with cooking dinner in the kitchen	He cut tomatoes into pieces.
19	4	Help with cooking dinner in the kitchen	He sliced the mushrooms.
20	5	Help with cooking dinner in the kitchen	He washed some cabbages under running water.
21	1	Clean the dining room after dinner	He took the plates to the kitchen.
22	2	Clean the dining room after dinner	He collected the utensils.
23	3	Clean the dining room after dinner	He scraped the crumbs off the table and into the trash can.
24	4	Clean the dining room after dinner	He put the chairs back to where they were.
25	5	Clean the dining room after dinner	He brought the sauce bottles back to the kitchen.
ENDING	ENDING	ENDING	Jim went back home at around eight, feeling tired but also very relaxed.

B. Sample Fine-level Semantic (FS) Narrative



Fine-level Index	Sub-fine-level Index	Coarse-level Event Label	Fine-level Event Sentence
OPENING	OPENING	OPENING	Mason decided to visit the zoo during the weekend.
1	1	Get a free souvenir tattoo	He walked towards the tattoo booth.
2	2	Get a free souvenir tattoo	He chose a pattern from many different choices.
3	3	Get a free souvenir tattoo	He placed the tattoo sticker onto his arm.
4	4	Get a free souvenir tattoo	With the help of the staff, he wet the paper backing with a cloth.
5	5	Get a free souvenir tattoo	He peeled away the paper backing and got a cute tattoo on his arm.
6	1	Watch a sea lion show	He took the ticket out of his pocket.
7	2	Watch a sea lion show	He showed his ticket to the staff at the gate.
8	3	Watch a sea lion show	He found a close seat for the performance.
9	4	Watch a sea lion show	The show began, and he watched the sea lion doing tricks.
10	5	Watch a sea lion show	He stood up with other audience members to give an ovation.
11	1	Visit the snack cart	He waited in the line of customers for several minutes.
12	2	Visit the snack cart	When it was his turn, he told the cart owner the type of snack he wanted.
13	3	Visit the snack cart	He paid with cash and received the food.
14	4	Visit the snack cart	He tasted the food at a bench nearby.
15	5	Visit the snack cart	He threw the empty food box away.
16	1	Feed pigeons at a square	He checked the notice to see if it was allowed to feed pigeons there.
17	2	Feed pigeons at a square	After getting confirmation, he went to buy a bag of seeds from the cart.
18	3	Feed pigeons at a square	He opened the bag and put some grains in his hand.
19	4	Feed pigeons at a square	He reached out his palm and waited for the pigeons to come.
20	5	Feed pigeons at a square	He felt the peck of a pigeon on his palm.
21	1	Participate in an animal knowledge quiz	He listened carefully to the question.
22	2	Participate in an animal knowledge quiz	He raised his hand immediately since he knew the answer.
23	3	Participate in an animal knowledge quiz	He was soon selected by the host of the quiz.
24	4	Participate in an animal knowledge quiz	He stood up and spoke the answer loudly.
25	5	Participate in an animal knowledge quiz	His answer was correct, and he got a special pin as a prize.
ENDING	ENDING	ENDING	Mason left the zoo in the evening. He was surprised that going to a zoo alone could also be very fun.

Figure 1. Example stimuli used in Experiment 1-3. Each narrative had 5 coarse-level event labels and 27 sentences, including one opening sentence at the beginning, one ending sentence at the end, and 25 fine-level event sentences in the middle. (A) Coarse-level Semantic (CS) Narrative: There were semantic order constraints on the order of the five coarse-level event labels, but there were no semantic order constraints among the five fine-level event sentences within each coarse-level event. (B) Fine-level Semantic (FS) Narrative: There were no semantic order constraints on the order of the five coarse-level event labels, but there were semantic order constraints on the five fine-level event sentences within each coarse-level event. Black arrows in the figure indicate the direction of semantic order constraints. Sentences pairs highlighted in orange indicate across-event pairs, and sentence pairs highlighted in blue indicate within-event pairs that were probed in the test phase.



He took out his hoodie from the closet.

He checked his aunt's address on the phone.

He made sure his gift for his aunt was already wrapped.

He sprayed some cologne on his neck.

He brushed his hair with a comb.

He put the car sun visor down so that he could see better.

He turned on the wipers to clean the windshield.

Please select the event that occurred first in the story.

He checked his aunt's address on his phone.

He brushed his hair with a comb.

On a scale of 1-100, what's your confidence for the previous order judgment?

On a scale of 1-10, how far apart in time were the two events presented in the story?

Recall what happened in the story according to the order it was presented...

E.g., At home, he sprayed perfume on his neck, brushed hair, checked aunt's address. In the car, he turned on the wipers......

Figure 2. Schematic of experimental procedure. (A) Encoding phase in Experiment 1-3. Participants read each narrative sentence-by-sentence in a self-paced format. Fine-level event sentence changed every one screen, and coarse-level event labels changed every five screens. In Experiment 2 and 3, apart from reading the contents on the screen, participants also heard audio of each fine-level event sentence. (B) Test phase in Experiment 1 & 2. For each fine-level event sentence pair, participants performed recency judgment task, rated their confidence for the recency judgment task, and then rated their perceived temporal distance between the two sentences. (C) Test phase in Experiment 3. Participants performed a serial recall task for the each of the narratives they read before the delay phase.

	Experiment 1	Experiment 2	Experiment 3
Sample Size (after data exclusion)	N = 40	N = 27	N = 32
Encoding Stimuli	10 narratives (5 CS + 5 FS)	10 narratives (5 CS + 5 FS)	6 narratives (3 CS + 3 FS)
Encoding Task	Self-paced reading	Self-paced reading with audio	Self-paced reading with audio
Delay Length	~ 2.5 minutes	~ 20 minutes	~ 2.5 minutes
Retrieval Task	(a) Recency Judgment (b) Recency Judgment Confidence (c) Temporal Distance Rating	(a) Recency Judgment (b) Recency Judgment Confidence (c) Temporal Distance Rating	(a) Serial Recall

Table 1. Comparison of Experiment 1-3 Method

Chapter 2: Experiment 1

In Experiment 1, we aimed to test the effect of event boundaries on temporal order memory in a narrative reading paradigm, where either coarse-level or fine-level semantic order knowledge could be utilized to facilitate with temporal order reconstruction. We hypothesized that people would perceive coarse-level event changes as event boundaries, which would lead to increased reading time when spatial changes happen (Radvansky et al., 2001; Zwaan & Radvansky, 1998). As a result, during retrieval, people would also rate fine-level events that span across event boundaries as farther apart in time comparing to fine-level events in the same coarse-level event, even though they were separated by the same number of sentences in between (DuBrow & Davachi, 2013; Ezzyat & Davachi, 2014). Critically, we hypothesized that the effect of event boundaries on recency judgment would depend on where semantic order knowledge could be utilized: In CS narratives, where coarse-level events were governed by semantic order constraints, it would be easier to discriminate the temporal order of two fine-level events from two different coarse-level events (*CS_across* pairs). But in FS narratives, where fine-level events within the same coarse-level event were governed by semantic order constraints, it would be easier to discriminate the temporal order of two fine-level events from the same coarse-level event (*FS_within* pairs). We predicted that the recency judgment accuracy of these two conditions would be better than the two conditions without semantic order knowledge facilitation (*CS_within* and *FS_across* pairs). In addition, we believed that because event models are formed hierarchically during reading comprehension, the order of coarse-level events would serve as an important information for recency judgment between fine-level events. On that view, we hypothesized that in the two conditions without semantic order knowledge facilitation, people would have more confidence in judging the order for FS_across pairs than CS_within pairs, because the order of coarse-level events they remember as a form of episodic memory might serve as an extra source of information.

The Institutional Review Board at Washington University in St. Louis approved the studies. We preregistered all the three experiments reported in this manuscript on Open Science Framework (https://osf.io). The preregistrations are available through the following URLs: Experiment 1 (https://osf.io/42d6p), Experiment 2 (https://osf.io/8wxhf), and Experiment 3 (https://osf.io/6j4k9).

2.1 Method

2.1.1 Participants

Seventy-four undergraduate students at Washington University in St. Louis participated in the experiment through the SONA subject pool in exchange for course credits. The mean age of the participants was 20.12 years (min = 18, max = 23, SD = 1.15). Fifty-two participants identified as female, twenty-one identified as male, and one identified as intersex, nonbinary, or other. Informed consent was obtained from all participants prior to the start of data collection.

We determined the sample size by performing a bootstrapped power analysis using a pilot sample (n = 25, 8 dropped based on the exclusion criteria, totaling n = 17) collected using Washington University SONA subject pool. We randomly sampled test data from the pilot sample with replacement to create 1000 new datasets with sample size ranging from 15 to 45 in steps of 5. We then ran a mixed effects logistic regression model to predict the accuracy of each recency judgment question (correct/incorrect) as a function of narrative type (CS/FS) and pair type (within/across), and their interaction. We calculated the proportion of simulations at each sample size that yielded difference estimate in the hypothesized direction and a p-value < 0.05 for the smallest effect we hypothesized. The power analysis showed that we needed at least 30 participants to achieve a statistical power of 80%. We recruited more participants than what was indicated by the power analysis, because we expected excluding some participants based on the criteria described below.

2.1.2 Materials

The design of the stimuli is depicted in Figure 1. To construct the stimulus set, we wrote ten narratives about everyday activities. Each narrative is consisted of 27 sentences, including one opening sentence at the beginning, one ending sentence at the end, and 25 fine-level event sentences in the middle. Each fine-level sentence was accompanied by a coarse-level event label that described its context, and the label changed after every five fine-level event sentences, totaling 5 coarse-level event labels in each narrative. To effectively create event boundaries at coarse-level event transitions, whenever a coarse-level event label changed to a new one, there was a spatial shift happening in the narrative (e.g., from “at home” to “in the car”).

The ten narratives were divided into two conditions, with five narratives in each condition: For Coarse-level Semantic (CS) narratives, all the coarse-level event labels in the narrative had a common-knowledge temporal order constraint (e.g., *preparing at home, and then driving the car to his aunt’s house, and then get greeted in his aunt’s living room*, etc.), whereas each set of fine-level events contained in each coarse-level event had no semantic order constraint (e.g. *take out hoodie, check aunt’s address, make sure the gift is wrapped*, etc.). For Fine-level Semantic (FS) narratives, all the coarse-level events had no semantic order constraint (e.g. *get a souvenir tattoo, watch a sea lion show, visit a snack cart*), whereas each set of fine-level events contained in each coarse-level event had a common-knowledge temporal order constraint (e.g. *wait in the line, then order the food, and then pay with cash*).

We verified the construction of our stimuli with an online norming study using a separate sample. The norming study was conducted to make sure that (1) the sets of coarse-level events in CS narratives and the sets of fine-level events in FS narratives have semantic order constraints that people generally agreed on, and (2) the sets of coarse-level events in FS narratives and the sets of fine-level events in CS narratives do not have semantic order constraints that people generally agreed on, and can instead be arranged in any order to make sense. Details of the norming study are reported in Supplemental Methods.

2.1.3 Procedure and Design

All the experiments reported in this manuscript were programmed using jsPsych (https://www.jspsych.org/7.3/) and were hosted online using Cognition (https://www.cognition.run/).

In Experiment 1, each participant completed 10 task runs, and each run corresponded to one of the 10 narrative stimuli (5 CS narratives and 5 FS narratives). Each run was consisted of an encoding phase, a delay phase, and a test phase.

In the encoding phase, participants were instructed to read a narrative sentence by sentence in a self-paced format. On each screen, they would encounter one sentence describing an activity ("fine-level event sentence," e.g., *"He took out his hoodie from the closet."*), as well as a label that gives them the context of the current activity ("coarse-level event label," e.g., *"Prepare at home"*). They were instructed to remember activities that happened in the story and click on the "Next" button to proceed to the next screen after finishing reading current contents. Figure 2(A) showed an illustration of the encoding phase.

After finishing each encoding phase, participants entered a delay phase in which they solved 40 math questions for approximately 150 seconds. They were asked to choose whether a given math question (e.g., 3*5-7) produced an odd or even result, and feedback ("Correct!" or "Wrong!") was given immediately after each answer. Participants were asked to not rush over the questions and try their best to maximize the accuracy.

After the delay phase, participants entered a test phase in which they were told to answer some questions based on the story. Each test phase consisted of eight trials. On each trial, they were presented with a pair of two fine-level event sentences selected from the narrative, and were asked to (1) make a recency judgment ("Please select the event that occurred first in the story"), (2) indicate their confidence for the recency judgment (On a scale of 1-100, what's your confidence for the previous order judgment?), and (3) give their rating of perceived temporal distance between these two sentences (On a scale of 1-10, how far apart in time were the two events presented in the story?). For each pair of fine-level sentences being probed, they were either within-event pairs (i.e., two sentences studied at the second and the fifth positions of the same coarse-level event), or across-event pairs (i.e., one sentence studied at the third position of a given coarse-level event, and another sentence studied at the first position in the next adjacent coarse-level event). Critically, both within-event sentence pairs and across-event sentence pairs were lag-2 pairs that were separated by two sentences in between during encoding time. Depending on whether the narrative being encoded in the current run was a Coarse-level Semantic (CS) or Fine-level Semantic (FS) narrative, a given test pair could be one of four types:

within-event pairs from Coarse-level Semantic narratives (CS_within), across-event pairs from Coarse-level Semantic narratives (CS_across), within-event pairs from Fine-level Semantic narratives (FS_within), across-event pairs from Fine-level Semantic narratives (FS_across). (Note that we conducted a separate semantic similarity analysis to make sure test pairs in different conditions do not differ significantly in terms of the semantic similarity between two sentences, in order to rule out semantic similarity as a confounding variable that might influences the memorability of the temporal relationship between sentences. Details of the semantic similarity analysis are reported in Supplemental Methods.) All the test questions were self-paced, and participants were instructed to use the "Next" button to proceed to the next question after finishing each question. Figure 2(B) showed an illustration of the testing phase.

The experiment was conducted using a 2 (Narrative Type: CS vs. FS) × 2 (Pair Type: Within vs. Across) within-subject design. Both Narrative Type and Pair Type were within-subject variables. The presentation order of ten narratives was randomized for each participant, and the order of event pairs being tested for a given narrative were randomized for each run.

2.1.4 Data Preparation

The final sample included responses from forty participants. Thirty-four Participants were excluded from the data analysis based on the exclusion criteria that we preregistered. Two were excluded for reporting experiencing technical problems during the online experiment, eight were excluded for having less than 75% accuracy for math questions during the delay phase, fourteen were excluded for having response times greater than 40000 ms for more than five encoding or test trials, and ten were excluded for having response times less than 300 ms for more than five encoding or test trials.

In addition, we excluded outlier trials based on the exclusion criteria that we preregistered. For encoding trials, we excluded all trials that had response times less than 300 ms, or more than 3SDs above the mean response time of sentences that are not the opening (first) or the ending (last) sentences of each narrative (0.2% of the data). For test trials, we excluded all trials that had response times less than 300 ms or greater than 3SDs above the mean reaction time for recency judgment trials (1.8% of the data), recency judgment confidence trials (1.3% of the data), and temporal distance rating trials (0.9% of the data).

We excluded the reading time data for the opening (first) and ending (last) sentence for each narrative. Reading time data were log-transformed to correct for skewness. In addition, we transformed recency judgment confidence (on a scale of 1-100) into a binary confidence group variable by coding confidence scores greater than 90 as in the "High Confidence" group, and coding confidence scores less than 90 as in the "Low Confidence" group. This was determined based on the distribution of confidence data in the pilot dataset, in which confidence = 100 was the mode, occurring in 40% of all confidence trials.

2.1.5 Analyses

We conducted data analyses using R. We estimated Linear Mixed-Effects Models using the lmer function, and Generalized Linear Mixed-Effects Models using the glmer function from the lme4 package (Bates et al., 2015). We started by fitting the "maximal model" (Barr et al., 2013) that included random slopes of all predictors. We removed a single random effect each time, and used a likelihood ratio test to compare the reduced model with the more complex model. We retained the most parsimonious model that did not differ significantly from the more complex model on model fit for each analysis (Bates et al., 2018).

2.2 Results

2.2.1 Reading Time

Based on previous studies showing that reading time was increased at event boundaries (Radvansky et al., 2001; Zwaan & Radvansky, 1998), we predicted that reading time for boundary sentences (i.e. the first sentence in each coarse-level event) would be longer than non-boundary sentences (i.e. the second to fifth sentences in each coarse-level event), when controlling for whether the sentence came from the first coarse-level event of each narrative, and controlling for narrative type (CS vs. FS). In addition, we predicted that in the first coarse-level event in a narrative, the extent to which boundary sentences required longer encoding time than other sentences would be larger than in other coarse-level events.

Mean reading time for sentences during the encoding phase (excluding opening and ending sentences) was 2131 ms (*SD* = 3986 ms). To satisfy model assumptions, we log-transformed reading time before entering it into the regression model. We predicted the log-transformed reading time of each sentence using a linear mixed-effects model, with fixed effects of fine-level position type (boundary vs. non-boundary sentence), coarse-level position type (first coarse event vs. other coarse event), narrative type (CS vs. FS narrative), and the interaction between coarse-level position type and fine-level position type. For the fine-level position type predictor, the boundary sentence condition was the reference condition. For the coarse-level position type predictor, the first coarse event condition was the reference condition. For the narrative type predictor, the CS narrative condition was the reference condition. The model comparison and random effects selection process led us to retain the random slope of narrative type on participant and random intercept of narrative as random effects. In Wilkinson notation, the model can be described as follows*: Log-transformed Reading Time ~ Fine-level Position Type + Coarse-level Position Type + Narrative Type + Coarse-level Position Type:Fine-level Position Type + (Narrative Type | Participant) + (1| Narrative)*.

Figure 3(A) shows estimated log-transformed reading time from the model by conditions. The linear mixed-effects model showed a significant main effect of fine-level position type, $F(1, 9891.01) = 155.14$, $p < .001$, a significant main effect of coarse-level position type, $F(1, 9891.03) = 116.28$, $p < .001$, no significant main effect of narrative type, $F(1, 11.62) = .003$, $p = 0.96$, and a significant interaction between fine-level position type and coarse-level position type, $F(1, 9891.03) = 19.40$, $p < .001$. The significant main effect of fine-level position type confirmed our hypothesis that boundary sentences (M = 7.55, SE = 0.06) required longer reading time (log-transformed) than non-boundary sentences (M = 7.32, SE = 0.06) after controlling for other factors, which suggests that participants perceived coarse-level event change as event boundaries during encoding. We probed the fine-level position type × coarse-level position type interaction with planned contrasts and found that the effect of being a boundary sentence on reading time was stronger in the first coarse-level event in a narrative ($B = 0.31$, $SE = 0.03$, $z = 9.42$, $p < .001$) than in other coarse-level events ($B = 0.15$, $SE = 0.02$, $z = 9.01$, $p < .001$), when controlling for narrative type.

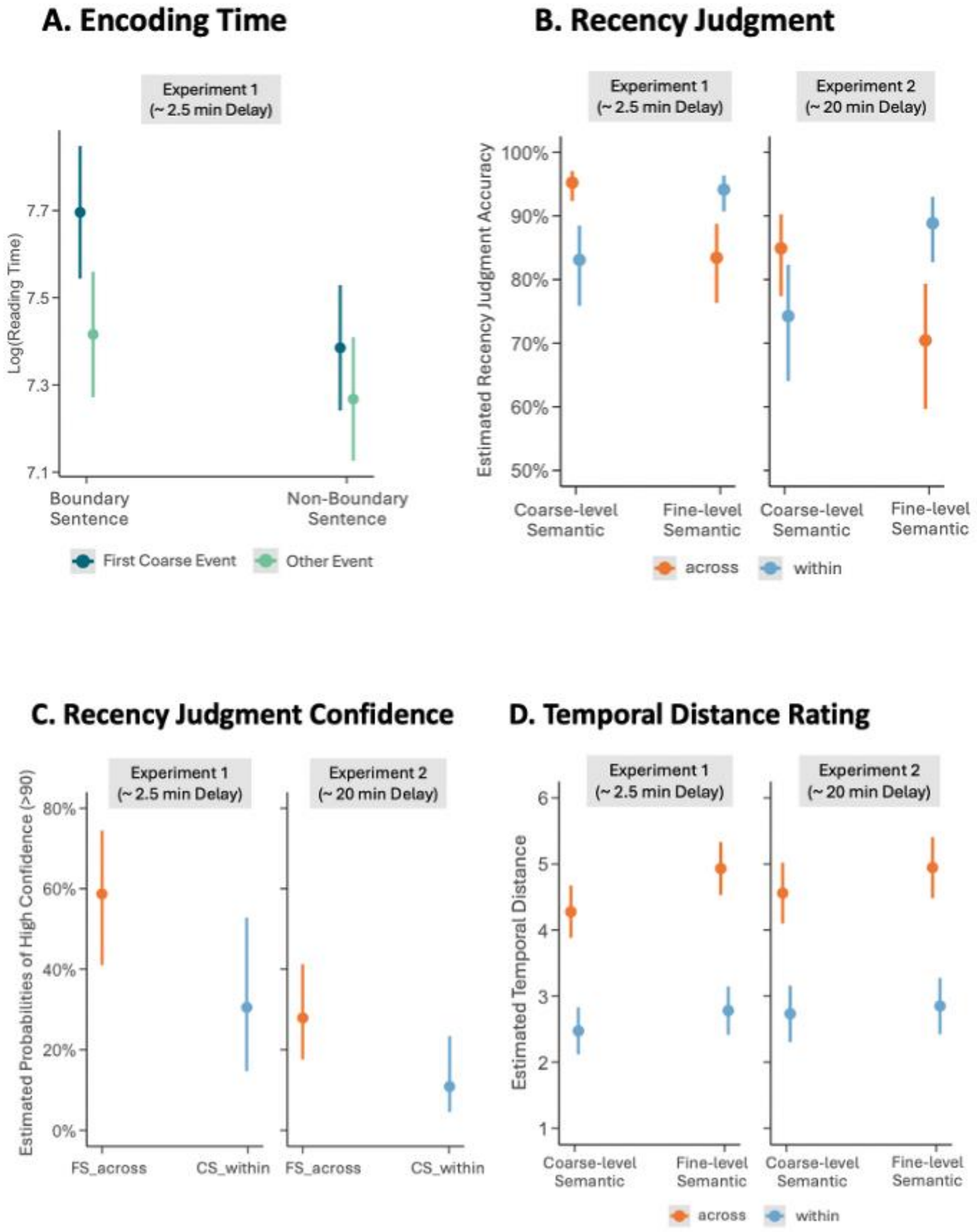


Figure 3. Experiment 1 and 2 results. (A) Estimated log-transformed reading time as a function of fine-level position type and coarse level position type from Experiment 1. (B) Estimated accuracy of recency judgment as a function of narrative type and fine-level event pair type from Experiment 1 (left) and 2 (right). (C) Estimated recency judgment confidence for FS_across and CS_within conditions from Experiment 1 (left) and 2 (right). (D) Estimated temporal distance rating as a function of narrative type and fine-level event pair type from Experiment 1 (left) and 2 (right). Error bars represent 95% confidence intervals.

2.2.2 Recency Judgment Accuracy

We hypothesized that whether the presence of event boundary impairs or facilitates temporal order memory depends on whether semantic order knowledge facilitation applies to coarse-level events or fine-level events. When there is semantic knowledge that helps people infer the temporal order between coarse-level events, temporal order memory of across-event pairs (CS_across pairs) will be better than when there is no semantic knowledge facilitation (FS_across and CS_within pairs); similarly, when there is semantic knowledge facilitation among fine-level events, temporal order memory of within-event pairs (FS_within pairs) will be better than when there is no semantic knowledge facilitation (CS_within and FS_across pairs).

Across all trials, mean recency judgment accuracy was .86 (SD = .35). We predicted whether a given recency judgment trial was correct or not using a logistic mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative), fine-level event pair type (across vs. within), and the interaction between narrative type and fine-level event pair type. For the narrative type predictor, the CS narrative condition was the reference condition. For the fine-level event pair type predictor, the across-event pair condition was the reference condition. The model comparison and random effects selection process led us to retain the random intercept of subject and event pairs as random effects. In Wilkinson notation, the model can be described as follows*: Recency Judgment Result (0/1) ~ Narrative Type + Fine-level Event Pair Type + Narrative Type:Fine-level Event Pair Type + (1 | Participant) + (1| Event Pair).*

Figure 3(B) shows estimated probabilities of correct recency judgment from the model by conditions. The logistic mixed-effects model showed a significant interaction between narrative type and fine-level event pair type, $X^2(1) = 36.45$, $p < .001$, but there was no significant main effect of narrative type, $X^2(1) = .20$, $p = .65$, and no significant main effect of event pair type, $X^2(1) = .32$, $p = .57$. We further probed the interaction and tested the four hypothesized pairwise contrasts: We found that CS_across pairs had better accuracy than both FS_across pairs (B = 1.38, SE = 0.30, $p < .001$) and CS_within pairs (B = 1.40, SE = 0.30, $p < .001$), and FS_within pairs had better accuracy than both CS_within pairs (B = 1.19, SE = 0.30, $p < .001$) and FS_across pairs (B = 1.16, SE = 0.30, $p < .001$). This confirmed our hypothesis that semantic order knowledge could be used on either level to improve the accuracy of recency judgment.

2.2.3 Recency Judgment Confidence

We hypothesized that due to the hierarchical event structure in the narratives, coarse-level episodic information would improve people's confidence on inferring the temporal order between fine-level event pairs. Specifically, when there was no influence of semantic knowledge on temporal order, temporal order memory confidence for across-event pairs in fine-level semantic narratives (FS_across pairs) would be better than within-event pairs in coarse-level semantic narratives (CS_within pairs), controlling for whether the trial was correct.

Across all trials, the mean recency judgment confidence was 76.53 (SD = 28.16) on a scale of 1 to 100. The confidence rating distribution was highly left-skewed, with 48% of the scores higher than 90. Therefore, we binarized the confidence variable used 90 as a cutoff, coding confidence scores greater than 90 as "High Confidence" and confidence score less than 90 as "Low Confidence." In addition, because the key comparison was between the recency judgment confidence of FS_across pairs and CS_within pairs, we decided to dummy code the fine-level event pair type variable (with FS_across as the reference group, and three dummy variables for FS_within, CS_across, and CS_within). We predicted recency judgment confidence (high/low) of a given sentence pair using a logistic mixed-effects model, with the fixed effects of recency judgment result (0/1), and three dummy variables FS_within, CS_across, and CS_within. For the recency judgment accuracy predictor, incorrect (0) was the reference condition. The model comparison and random effects selection process led us to retain the slope of three dummy variables of FS_within, CS_across, and CS_within on subject and random intercepts of event pairs as random effects. In Wilkinson notation, the model can be described as follows*: Recency Judgment Confidence (High/Low) ~ Recency Judgment Result (0/1) + FS_within + CS_across + CS_within + (FS_within + CS_across + CS_within | Participant) + (1| Event Pair)*.

Figure 3(C) shows the estimated probabilities of rating a given recency judgment trial as highly confident (higher than 90) from the model by conditions. The logistic mixed-effects model showed a significant main effect of recency judgment result, $X^2(1) = 73.98$, $p < .001$, a significant main effect of event pair type CS_within (compared to the reference group FS_across), $B = -1.17$, $X^2(1) = 14.42$, $p < .001$, and a significant main effect of event pair type FS_within (compared to the reference group FS_across), $B = 1.15$, $X^2(1) = 13.71$, $p < .001$. There was no significant effect of event pair type CS_across (compared to the reference group FS_across), $B = 0.352$, $X^2(1) = 1.32$, $p = 0.25$. The significant main effect of CS_within confirmed our hypothesis that participants had higher confidence for recency judgment for FS_across pairs (M = -.20, SE = .37) than for CS_within pairs (M = -1.37, SE = .49), after controlling for accuracy. This suggested that when there was no semantic order knowledge facilitation, participants had higher confidence for the relative recency of event pairs coming from two different coarse-level events (FS_across) than for event pairs coming from the same coarse-level events (CS_within), controlling for the accuracy of the judgment.

2.2.4 Temporal Distance Rating

Based on previous studies showing temporal distance rating inflation due to the presence of event boundaries (DuBrow & Davachi, 2013; Ezzyat & Davachi, 2014), we hypothesized that the temporal distance between two fine-level events would be rated as farther if they spanned across an event boundary, and this effect would not be influenced by where semantic order constraints existed in the narratives.

Across all trials, the mean temporal distance rating was 3.61 (SD = 2.00) on a scale of 1 to 7. We predicted the temporal distance rating of a given sentence pair using a linear mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative), fine-level event pair type (across vs. within), and the interaction between narrative type and fine-level event pair type. For the narrative type predictor, the CS narrative condition was the reference condition. For the fine-level event pair type predictor, the across-event pair condition was the reference condition. The model comparison and random effects selection process led us to retain the random slopes of narrative type, fine-level event pair, and their interaction on subject, and random intercepts of event pairs as random effects. In Wilkinson notation, the model can be described as follows*:* *Temporal Distance Rating ~ Narrative Type + Fine-level Event Pair Type + Narrative Type:Fine-level Event Pair Type + (Narrative Type + Fine-level Event Pair Type + Narrative Type:Fine-level Event Pair Type | Participant) + (1| Event Pair)*.

Figure 3(D) shows estimated value of temporal distance rating from the model by condition. The linear mixed-effects model showed a significant main effect of narrative type, $F(1, 84.29) = 13.79$, $p < .001$, a significant main effect of fine-level event pair type, $F(1, 92.19) = 168.68$, $p < .001$. There was no significant interaction between narrative type and fine-level event pair type, *F*(1, 79.09) = 1.89, *p* = .17. To highlight, the significant main effect of fine-level event pair type confirmed our hypothesis that across-event pairs (M = 4.60, SE = 0.18) were rated as farther away than within-event pairs (M = 2.63, SE = 0.16) after controlling for other factors, which suggested that the presence of event boundary inflated temporal distance rating. We further probed the interaction and tested the four hypothesized pairwise contrasts: We found that CS_across pairs were perceived as more temporally distant than both CS_within pairs (*B* = 1.80, *SE* = 0.19, *p* < .001) and FS_within pairs (*B* = 1.50, *SE* = 0.20, *p* < .001), and FS_across pairs were perceived as more temporally distant than both CS_within pairs (*B* = 2.46, *SE* = 0.20, *p* < .001) and FS_within pairs (*B* = 2.15, *SE* = 0.20, *p* < .001).

2.3 Discussion

Our primary goal in Experiment 1 was to examine whether semantic order knowledge could interact with hierarchical event structure to facilitate temporal order memory. First, we showed that changes in coarse-level events in our narratives were reliably perceived by the participants as event boundaries. This was supported by two converging pieces of evidence: (1) Increased reading time for boundary sentences comparing to non-boundary sentences during encoding, and (2) inflated temporal distance judgment for across-event sentence pairs comparing to within-event pairs during retrieval. Second, for recency judgments, we found that participants could use semantic constraints at either level to facilitate temporal order memory. This was supported by our finding that when semantic order knowledge was present at coarse-level (in CS narratives), across-event recency judgment is more accurate than within-event recency judgment; when semantic order knowledge was present at fine-level (in FS narratives), within-event recency judgment is more accurate than across-event recency judgment. Specifically, the advantage that FS_within pairs received suggested that semantic order constraint on the fine-level can be used as a source of information for fine-level recency judgment. The advantage that CS_across pairs received suggested that semantic order constraint on the coarse-level can also be used as a source of information for fine-level recency judgment, suggesting that participants construct and store event models hierarchically during reading comprehension. Together, these finding suggested that the role event boundaries play in shaping temporal order memory task performance depends on how semantic order knowledge can be leveraged to facilitate memory reconstruction.

In addition, we compared the confidence ratings for recency judgment between the two conditions when semantic order knowledge could not be leveraged and found that when controlling for whether the outcome is correct, participants were more confident about recency judgments outcome in FS_across condition than in CS_within condition. One important difference between these two conditions is that the two fine-level event sentences in FS_across pairs came from two different coarse-level events, but the two sentences in CS_within pairs came from the same coarse-level event. Semantic order knowledge could not be used in either of these two conditions, because there were no semantic order constraints among coarse-level events in FS narratives or among fine-level events in the CS narratives. The fact that participants were more confident about FS_across pairs again suggested people encode hierarchical relationship among events during reading comprehension, and they might be using episodic coarse-level association that was formed during encoding as a source of information for fine-level recency judgment.

One limitation of Experiment 1 was that following the design of most picture-list learning paradigms, we used a relatively short delay task (~ 2.5 minutes) between encoding and retrieval.

With meaningful narratives as stimuli, we were also curious about whether the patterns we observed would remain the same after a longer delay period, which would make it more convincing to make generalizations about how temporal order relationship is reconstructed from long-term memory.

Chapter 3: Experiment 2

In Experiment 2, we aimed to replicate some of our key findings in Experiment 1 using a longer delay procedure. Instead of testing each narrative after its own encoding block followed by a short distraction task, we decided to have participants encode all the ten narratives at once, and then receive tests regarding these narratives according to the order they were encoded. In this way, we created a natural delay of about 20 minutes between the encoding and retrieval of each narrative and filled the delay period with learning and testing on other stimuli. Thus, we sought to observe if the patterns we observed about temporal order memory would hold true after more chances of interference and memory decay that resembled memory retention in the real world.

In order to increase participants' engagement during a long encoding period, we played audio for sentences in each narrative during encoding, and participants could only proceed to the next screen after the audio of each sentence finished playing. This led us to drop the hypothesis regarding longer boundary sentence reading time from Experiment 1, because the length of audio play might constrain self-paced reading time. Apart from this hypothesis, we predicted that we would find the same pattern in recency judgment accuracy, recency judgment confidence, and temporal distance rating as in Experiment 1 using a longer retention interval in the current experiment.

3.1 Method

3.1.1 Participants

Thirty-eight undergraduate students at Washington University in St. Louis participated in the experiment through the SONA subject pool in exchange for course credits. The mean age of the participants was 19.47 years (min = 18, max = 22, SD = 1.20). Twenty-six participants identified as female, and twelve identified as male. Informed consent was obtained from all participants prior to the start of data collection.

We determined the sample size by performing a bootstrapped power analysis using a pilot sample (n = 15, 5 dropped based on the exclusion criteria, totaling n = 10) collected using Prolific (https://www.prolific.com/). We randomly sampled test data from the pilot sample with replacement to create 1000 new datasets with sample size ranging from 15 to 35 in steps of 5. We then ran a mixed effects logistic regression model to predict the accuracy of each recency judgment question (correct/incorrect) as a function of narrative type (CS/FS) and pair type (within/across), and their interaction. We calculated the proportion of simulations at each sample size that yielded difference estimate in the hypothesized direction and a p-value < 0.05 for the smallest effect we hypothesized. The power analysis showed that we needed at least 25 participants to achieve a statistical power of 90%. We recruited more participants than what was indicated by the power analysis, because we expected excluding certain participants based on the exclusion criteria.

3.1.2 Materials

The stimuli were the same ten narratives about everyday activities used in Experiment 1. Additionally, we generated audios for sentences in all the narratives using an online AI-based text-to-speech tool, ElevenLabs (https://elevenlabs.io/).

3.1.3 Procedure and Design

In order to create a 20-minute delay between encoding and test for each narrative in Experiment 2, we decided to have a long encoding block for encoding all ten narratives one by one sequentially, followed by a long test block for testing all narratives in the same order.

In the encoding block, participants completed the encoding phases of all ten narratives consecutively in a randomized order. Same as in Experiment 1, they were instructed to read each narrative sentence by sentence in a self-paced format. The on-screen display of fine-level event sentences and coarse-level event labels were the same as in Experiment 1, except that participants also heard the audio of each fine-level event sentence as they read. They could only click on the "Next" button to proceed to the next sentence after the audio for the current sentence finished playing. This change was made to increase participants' engagement in the task, but we also acknowledged that it would constrain the encoding time for each sentence. Therefore, we dropped the hypotheses regarding how narrative structure influenced reading time for each sentence that were tested in the previous experiment. Figure 2(A) showed an illustration of the encoding phase.

Another change from Experiment 1 was that participants answered two reading comprehension questions immediately after reading each narrative. This change was also made to increase participants' engagement in the task. Each of these reading comprehension questions probed one specific detail in the narrative, and participants chose one option out of four options provided. Here is a sample reading comprehension question: *"What did Jim notice outside of his car window?"* (Correction option: *"The trees turning yellow."*) To avoid interference with the later test block, we carefully constructed these questions so that they never probe contents in the fine-level event sentences being tested in the recency judgment and distance rating tasks.

After encoding all ten narratives in the encoding block, participants entered the test block. They completed the test phase of each narrative according to their presentation order in the encoding block. Before entering each test phase, they received an instruction specifying the narrative being probed (e.g., *"The following questions are for the story about Jim visiting his aunt."*). Each test phase tested the same eight fine-level event sentence pairs as in Experiment 1. For each pair, they were asked to (1) make a recency judgment, (2) indicate their confidence for the recency judgment, and (3) give their rating of perceived temporal distance between these two sentences. Figure 2(B) showed an illustration of the testing phase.

As in Experiment 1, Experiment 2 was conducted using a 2 (Narrative Type: CS vs. FS) × 2 (Pair Type: Within vs. Across) within-subject design. Both narrative Type and pair type are within-subject variables. The presentation order of ten narratives (which is the same as their order of being tested) were randomized for each participant, and the order of event pairs being tested for a given narrative were randomized for each run.

3.1.4 Data Preparation

The final sample included responses from twenty-seven participants. Eleven Participants were excluded from the data analysis based on the exclusion criteria that we preregistered. One was excluded for reporting experiencing technical problems during the online experiment, three were excluded for having reaction time greater than 40000 ms for more than five encoding or test trials, four were excluded for having less than 60% accuracy for reading comprehension questions after encoding each narrative, and three were excluded for having less than 50% accuracy for fine-level recency judgment questions.

Additionally, we excluded outlier trials based on the exclusion criteria that we preregistered. For test trials, we excluded all trials that had reaction time less than 300 ms, or greater than 3SDs above the mean reaction time for recency judgment trials (1.6% of the data), recency judgment confidence trials (1.1% of the data), and temporal distance rating trials (1.3% of the data). Like in Experiment 1, we transformed recency judgment confidence (on a scale of 1-100) into a binary confidence group variable by coding confidence score greater than 90 as "High Confidence," and coding confidence score less than 90 as "Low Confidence."

3.1.5 Analyses

We conducted data analyses using R. We estimated Linear Mixed-Effects Models using the lmer function, and Generalized Linear Mixed-Effects Models using the glmer function from the lme4 package (Bates et al., 2015). We started by fitting the "maximal model" (Barr et al., 2013) that included random slopes of all predictors. We removed a single random effect each time, and used a likelihood ratio test to compare the reduced model with the more complex model. We retained the most parsimonious model that did not differ significantly from the more complex model on model fit for each analysis (Bates et al., 2018).

3.2 Results

3.2.1 Recency Judgment Accuracy

The results from Experiment 1 were replicated in Experiment 2 using the same analyses. Across all trials, mean accuracy on recency judgment was .76 (SD = 0.43). We predicted whether a given recency judgment trial was correct or not using a logistic mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative), fine-level event pair type (across vs. within), and the interaction between narrative type and fine-level event pair type. For the narrative type predictor, the CS narrative condition was the reference condition. For the fine-level event pair type predictor, the across-event pair condition was the reference condition. The model comparison and random effects selection process led us to retain the random intercepts of subject and event pairs as random effects. In Wilkinson notation, the model can be described as follows*: Recency Judgment Result (0/1) ~ Narrative Type + Fine-level Event Pair Type + Narrative Type:Fine-level Event Pair Type + (1 | Participant) + (1| Event Pair).*

Figure 3(B) shows estimated probabilities of correct recency judgment from the model by conditions. The logistic mixed-effects model showed a significant interaction between narrative type and fine-level event pair type, $X^2(1) = 18.13$, $p < .001$, but there was no significant main effect of narrative type, $X^2(1) = 0.13$, $p = 0.72$, and no significant main effect of event pair type, $X^2(1) = 1.50$, $p = 0.22$. We further probed the interaction and tested the four hypothesized pairwise contrasts: We found that CS_across pairs had better accuracy than both FS_across pairs (B = 0.86, SE = 0.31, $p = .005$) and CS_within pairs (B = 0.67, SE = 0.31, $p = .03$), and FS_within pairs had better accuracy than both CS_within pairs (B = 1.02, SE = 0.32, $p = .001$) and FS_across pairs (B = 1.21, SE = 0.31, p < .001). This confirmed our hypothesis that semantic order knowledge could be used at either level to facilitate recency judgment accuracy, even after a 20-minute delay.

3.2.2 Recency Judgment Confidence

Results in Experiment 1 were replicated in Experiment 2 using the same analyses. Across all trials, the mean recency judgment confidence was 61.55 (SD = 33.17) on a scale of 1 to 100. The confidence rating distribution was highly left-skewed, with 28% of the scores higher than

90. Therefore, we decided to use 90 as a cutoff, coding confidence score greater than 90 as "High Confidence," and confidence score less than 90 as in the "Low Confidence" group. In addition, since the key contrast we would like to compare was the recency judgment confidence between FS_across pairs and CS_within pairs, we decided to dummy code the fine-level event pair type variable (with FS_across as the reference group, and three dummy variables for FS_within, CS_across, and CS_within). We predicted the recency judgment confidence (high / low) of a given sentence pair using a logistic mixed-effects model, with the fixed effects of recency judgment result (0/1), and three dummy variables FS_within, CS_across, and CS_within. For the recency judgment accuracy predictor, incorrect (0) was the reference condition. The model comparison and random effects selection process led us to retain the random intercepts of subject and event pairs as random effects. In Wilkinson notation, the model can be described as follows*:* *Recency Judgment Confidence (High/Low) ~ Recency Judgment Result (0/1) + FS_within + CS_across + CS_within + (1 | Participant) + (1| Event Pair).*

Figure 3(C) shows the estimated probabilities of rating a given recency judgment trial as highly confident (> 90) from the model by conditions. The logistic mixed-effects model showed a significant main effect of recency judgment result, $X^2(1) = 52.73$, $p < .001$, a significant main effect of event pair type CS_within (compared to the reference group FS_across), $B = -1.13$, $X^2(1) = 10.63$, $p < .001$, and a significant main effect of event pair type FS_within (compared to the reference group FS_across), $B = 1.23$, $X^2(1) = 13.80$, p < .001. There was no significant effect of event pair type CS_across (compared to the reference group FS_across), $B = 0.12$, $X^2(1) = 0.13$, p = 0.72. To conclude, the significant main effect of CS_within confirmed our hypothesis that participants had higher confidence for recency judgment for FS_across pairs (M = -.197, SE = 0.37) than for CS_within pairs (M = -1.37, SE = 0.49), after controlling for accuracy. This suggested that when there is no semantic order knowledge facilitation, participants have higher confidence for the relative recency of event pairs coming from two different coarse-level events (FS_across) than for event pairs coming from the same coarse-level events (CS_within), controlling for the accuracy of the judgment.

3.2.3 Temporal Distance Rating

Results in Experiment 1 were replicated in Experiment 2 using the same analyses. Across all trials, the mean temporal distance rating was 3.77 (SD = 2.12) on a scale of 1 to 7. We predicted the temporal distance rating of a given sentence pair using a linear mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative), fine-level event pair type (across vs. within), and the interaction between narrative type and fine-level event pair type. For the narrative type predictor, the CS narrative condition was the reference condition. For the fine-level event pair type predictor, the across-event pair condition was the reference condition. The model comparison and random effects selection process led us to retain the random slopes of fine-level event pair type on subject and random intercepts of event pairs as random effects. In Wilkinson notation, the model can be described as follows*: Temporal Distance Rating ~ Narrative Type + Fine-level Event Pair Type + Narrative Type:Fine-level Event Pair Type + (Fine-level Event Pair Type | Participant) + (1| Event Pair)*.

Figure 3(D) shows estimated value of temporal distance rating from the model by conditions. The linear mixed-effects model showed a significant main effect of fine-level event pair type, $F(1, 62.11) = 84.38$, $p < .001$. There was no significant main effect of narrative type, $F(1, 75.97) = 2.60$, $p = .11$, and there is no significant interaction between narrative type and fine-level event pair type, $F(1, 75.97) = 0.75$, $p = .39$. To highlight, the significant main effect of fine-level event pair type confirmed our hypothesis that across-event pairs (M = 4.75, SE = 0.21) were rated as farther away than within-event pairs (M = 2.79, SE = 0.19) after controlling for other factors, which suggested that the presence of event boundary inflated temporal distance rating. We further probed the interaction and tested the four hypothesized pairwise contrasts: We found that CS_across pairs were perceived as more temporally distant than both CS_within pairs (*B* = 1.83, *SE* = 0.19, *p* < .001) and FS_within pairs (B = 1.71, SE = 0.26, *p* < .001), and FS_across pairs were perceived as more temporally distant than both CS_within pairs (B = 2.21, SE = 0.26, p < .001) and FS_within pairs (B = 2.10, SE = 0.26, *p* < .001).

3.3 Discussion

In Experiment 2, we aimed to replicate key findings in Experiment 1 using a longer delay. We increased the length of the delay period between encoding and retrieval from 2.5 minutes to about 20 minutes and added more potentially interfering information during the retention period. We ended up replicating all the patterns observed in Experiment 1: For recency judgment, we again confirmed the hypothesis that semantic order knowledge could be used at either coarse-level and fine-level to facilitate recency judgment accuracy. For recency judgment confidence, we found that when there was no semantic knowledge facilitation, coarse-level event information might serve as an extra source of information for recency judgment, and thus increased subjective confidence when controlling for accuracy. For temporal distance rating, we observed the stable pattern that when two fine-level events were separated by an event boundary, they were rated as temporally farther apart from each other. Together, these results suggested that coarse-level event membership and semantic order constraints might influence how fine-level events were organized in long-term memory. To directly investigate the temporal organization of people’s memory about events in the narratives, we decided to test serial recall performance in the next experiment.

Chapter 4: Experiment 3

In Experiments 1 and 2, we used recency judgment tasks to examine the influence of semantic order knowledge and event hierarchy on temporal order memory. However, one potential limitation of the task design was that since both test probes were presented to the participants to make recency judgment, participants were forced to make a choice even if they did not remember the temporal relationship between the two probes. Thus, in order to understand how semantic order knowledge and event structure influence the temporal organization of events in long-term memory, we tested participants' serial recall of the narratives in Experiment 3. We asked participants to type down what they remember about the narrative according to the order it was presented after a short delay (as in Experiment 1) and analyzed the structure of their recall response. First, we hypothesized that the number of fine-level events being recalled would be influenced by whether the narrative contained semantic order constraint on the fine-level to facilitate reconstruction. This was based on previous literature suggesting that people were more likely to recall more information if they were directly related to an underlying script or linked by causal structure (Lichtenstein & Brewer, 1980; Radvansky & Zacks, 2017; Lee & Chen, 2022). Since only FS narratives contained semantic order constraints on the fine-level, we predicted that participants would recall more fine-level events per narrative or per coarse-level event for FS narratives, comparing to CS narratives. Second, we hypothesized that the order of fine-level events being recalled would be influenced by semantic order constraints as well, based on previous studies suggesting that the temporal order of recalled goal-directed events was strongly influenced by underlying schema (Lichtenstein & Brewer, 1980; Bower et al., 1979; Brewer & Dupree, 1983). Therefore, we predicted that for FS narratives, fine-level events recalled within each coarse-level event would be more in-order comparing to for CS narratives. In addition, we asked how the order of immediate transitions in serial recall was influenced by fine-level semantic order constraints. We predicted that in CS narratives, if adjacent fine-level events being recalled came from the same coarse-level event (CS_within transition), they were more likely to be recalled in the wrong order, compared to when adjacent fine-level events being recalled came from the same coarse-level event in FS narrative (FS_across transition) or when adjacent fine-level events being recalled came from different coarse-level events in CS narrative (CS_across transition).

4.1 Method

4.1.1 Participants

Thirty-eight undergraduate students at Washington University in St. Louis participated in the experiment through the SONA subject pool in exchange for course credits. The mean age of the participants was 20.12 years (min = 18, max = 23, SD = 1.23). Twenty-nine participants identified as female, and nine identified as male. Informed consent was obtained from all participants prior to the start of data collection.

We determined the sample size by performing a bootstrapped power analysis using a pilot sample (n = 9, 1 dropped based on exclusion criteria, totaling n = 8). All the hypothesized effects were in the right direction in the pilot sample, and we decided to power our study based on the smallest effect we observed by running simulation-based power analysis using the package "Mixedpower" in R (Kumle et al., 2021). The power analysis showed that we need at least 30 participants to achieve the power of 90%.

4.1.2 Materials

The stimuli we used in Experiment 3 were six out of ten narratives about everyday activities used in the previous experiments, which contained three Coarse-level Semantic (CS) narratives and three Fine-level Semantic (FS) narratives.

4.1.3 Procedure and Design

In the experiment, each participant completed 6 task runs, and each run corresponded to one of the 6 narrative stimuli (3 Coarse-level Semantic and 3 Fine-level Semantic). Each run consisted of an encoding phase, a delay phase, and a test phase.

The encoding and the delay phase used the same procedure as in Experiment 1, with the following exception: During the encoding phase, in addition to reading the fine-level event sentences and coarse-level event labels on the screen, participants also heard the audio of each fine-level event sentence. They were instructed to remember each activity in the story in as much detail as possible, as if they were preparing to retell the story to a friend later. In the delay phase, they solved 40 math questions for approximately 150 seconds, as in Experiment 1.

After the delay phase, participants entered a test phase. They were told to type down everything they could remember from the story they read before doing math, according to the order it was presented, in as much detail as possible. They were given at least two minutes to type down their response before proceeding to the next run using the “Next” button. If they did not finish after two minutes, they were allowed to take extra time to finish their typing.

The presentation order of the six narratives were randomized for each participant.

4.1.4 Data Preparation

The final sample included responses from thirty-two participants. Six participants were excluded from the data analysis based on the exclusion criteria that we preregistered. Three were excluded for reporting experiencing technical problems during the online experiment, two were excluded for having less than 75% accuracy for math questions during the delay phase, and one were excluded for giving empty response for at least one recall trial.

Additionally, we excluded outlier recall trials based on the exclusion criteria that we preregistered. We excluded all trials that had reaction time > 3SDs above the mean reaction time of all recall trials (1% of the data).

To score recall responses, we compared each participants' typed recall with the original script for each narrative. For each fine-level event sentence in the original script of a narrative, two coders identified the key verbs and key objects in the sentence based on the situation it described. For example, for the fine-level event sentence *"he took out his hoodie from the closet,"* "*took out*" was identified as the key verb phrase, and "*hoodie*" and "*closet*" were identified as the key objects. We counted a fine-level event as being recalled if at least one of the key verbs or the key objects in the original sentence or their synonyms was mentioned in the recall protocol, and if the recalled event corresponded to the original event on a situational level (van Dijk & Kintsch, 1983).

For each of the fine-level events in the script that was mentioned in the recall, we recorded the event's ordinal position in the script following the order it was mentioned in the recall. Therefore, for each typed recall, we derived a vector of ordinal positions representing the fine-level events being recalled (Diamond & Levine, 2020) for the whole narrative ("fine-level order vector"), as well as the sub-vectors of ordinal positions representing the fine-level events recalled within each coarse-level event (“sub-fine-level order vector”). For each recall response made by each participant, there were one fine-level order vector and up to five (depending on the number of coarse-level events being recalled) sub-fine-level order vectors being extracted.

For each coarse-level event in the script, we counted it as being recalled if the coarse-level label was directly mentioned in the recall. If the coarse-level event was not directly mentioned, we counted it as being recalled if at least one fine-level event from it was coded as recalled (Tulving & Pearlstone, 1966). We recorded the ordinal position of each coarse-level event in the script following the order it was mentioned in the recall and derived a vector of ordinal positions representing the coarse-level events being recalled (“coarse-level order vector”). For each participant’s recall of one story, there was one coarse-level order vector extracted.

Based on the three types of recall vectors, we computed the number of fine-level events recalled per narrative (i.e. the length of fine-level order vector), the number of fine-level events recalled per coarse-level event (i.e. the length of sub-fine-level order vector), and the number of coarse-level events recalled per narrative (i.e. the length of coarse-level order vector). To quantify how much fine-level events recalled within each coarse-level event deviated from the correct order, we computed a deviance score by comparing each sub-fine-level order vector (recall_vector) with a correct version of this recall_vector being sorted in ascending order (correct_vector). To compute a deviance score, for each adjacent transition, we subtracted the transition lag in correct_vector by the transition lag in recall_vector, took the absolute value of the subtraction, and sum across all adjacent transitions. For example, if the sub-fine-level order vector was [5, 1, 3, 2], which meant that the participant first recalled the fifth fine-level event in the coarse-level event, then the first, then the third, and then the second, it was compared with the vector [1, 2, 3, 5], which arranged all the events recalled in the correct order, to yield a deviance score of 9 ($9 = |(2-1)-(1-5)| + |(3-2)-(3-1)| + |(5-3)-(2-3)|$). A recall in the correct order would yield a deviance score of 0, and larger deviance score would indicate a bigger deviance from the correct order. We then divided the deviation score by the number of transitions made in each recall vector (number of transitions = recall vector length – 1) to create a normalized deviation score that was independent on length of the recall, and use this normalized deviation score as the outcome variable of the regression model.

To quantify the direction of recall transitions, we categorized all immediate (lag = 0) transitions among fine-level events in each recall into within-event transitions (i.e. two adjacent fine-level events belonging to the same coarse-level event) and across-event transitions (i.e. two adjacent fine-level events belonging to two different coarse-level events), and determine if each transition is in the same order as in the narratives presented during encoding. For example, if the 5th fine-level event was recalled right after the 3rd fine-level event, we would count this as a correct forward transition, even though the 4th event was skipped in the recall. We coded each transition as 1 if it was in the correct order, and 0 otherwise.

Two raters (YD and DA) were trained on the scoring criteria described above using data from the pilot experiment (n = 8). For the data collected in Experiment 2, YD and DA each scored recall data from half of the participants. To calculate interrater reliability, YD scored the recall data from five participants that DA also scored. The interrater reliability was relatively high (mean Cohen's Kappa = 0.86), which justified our decision of having each rater coding half of the recall data.

4.2 Results

4.2.1 Number of Fine-level Events Recalled Per Narrative

We hypothesized that participants would recall more fine-level events per narrative for FS narratives than for CS narratives, because previous works have suggested that events with more causal connections to other events are more likely to be remembered during retrieval (Radvansky & Zacks, 2017; Lee & Chen, 2022). In FS narratives, since each set of five fine-level events has semantic order constraints, they should be more likely to be recalled than fine-level events without order constraints in CS narratives.

On average, each participant recalled 12.01 (SD = 5.54) out of 25 fine-level events in each narrative. We predicted the number of fine-level events recalled in each narrative using a linear mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative). For the narrative type predictor, the CS narrative condition was the reference condition. The model comparison and random effects selection process led us to retain the random intercepts of subject and narrative as random effects. In Wilkinson notation, the model can be described as follows*:* *Number of Fine-level Events Recalled Per Narrative ~ Narrative Type + (1 | Participant) + (1| Narrative)*.

Figure 4(A) shows estimated number of fine-level events recalled in each narrative from the model by conditions. Contrary to our hypothesis, the linear mixed-effects model did not show a significant main effect of narrative type, $F(1, 4.00) = 0.05$, $p = 0.84$. The Bayes Factor value indicated weak evidence in support of the null hypothesis, $BF_{10} = 0.48$. In conclusion, we did not find strong evidence suggesting that the number of fine-level events recalled per narrative differ significantly across FS and CS narratives.

4.2.2 Number of Fine-level Events Recalled Per Coarse-level Event

We hypothesized that participants would recall more fine-level events per coarse-level event for FS narratives than for CS narratives, based on previous works suggesting that events with more causal connections to other events are more likely to be remembered during retrieval (Radvansky & Zacks, 2017; Lee & Chen, 2022)

On average, each participant recalled 2.85 (SD = 1.25) out of 5 fine-level events per coarse-level event. We predicted the number of fine-level events recalled in each coarse-level event using a linear mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative). For the narrative type predictor, the CS narrative condition was the reference condition. The model comparison and random effects selection process led us to retain the random intercept of subject and narrative as random effects. In Wilkinson notation, the model can be described as follows*: Number of Fine-level Events Recalled Per Coarse-level Event ~ Narrative Type + (1 | Participant) + (1| Narrative)*.

Figure 4(B) shows estimated number of fine-level events recalled in each coarse-level event from the model by conditions. Contrary to our hypothesis, the linear mixed-effects model did not show a significant main effect of narrative type, $F(1, 4.00) = 0.89$, $p = 0.40$. The Bayes Factor value indicated weak evidence in support of the null hypothesis, $\mathrm{BF}_{10} = 0.85$. In conclusion, we did not find strong evidence suggesting that the number of fine-level events recalled per coarse-level event differ significantly across FS and CS narratives.

4.2.3 Order of Fine-level Events Recalled within Each Coarse-level Event

Next, we hypothesized that participants would recall fine-level events within each coarse-level event less in-order for CS narratives than for FS narratives, because comparing to FS narratives, there is no semantic order constraints among fine-level events in CS narratives.

On average, the mean normalized order deviance score for fine-level events recalled within each coarse-level event is 0.31 (SD = 0.85). We predicted the magnitude of normalized recall order deviance score using a linear mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative). For the narrative type predictor, the CS narrative condition was the reference condition. The model comparison and random effects selection process led us to retain the random intercept of subject and narrative as random effects. In Wilkinson notation, the model can be described as follows*: Number of Fine-level Events Recalled Per Narrative ~ Narrative Type + (1 | Participant) + (1| Narrative)*.

Figure 4(C) shows estimated normalized order deviance score for fine-level events recalled within each coarse-level event from the model by conditions. The linear mixed-effects model showed a significant main effect of narrative type, $F(1, 4.16) = 19.81$, $p = 0.01$. This confirmed our hypothesis that fine-level events within each coarse-level event were recalled less in-order for CS narratives (M = 0.56, SE = 0.08) than for FS narratives (M = 0.04, SE = 0.08).

4.2.4 Forward Transition Probability between Adjacent Fine-level Events Recalled

We hypothesized that there would be lower probability of making transitions in the correct order if the adjacent fine-level events recalled were from the same coarse-level event in CS narratives. Specifically, correct forward within-event transition in CS narratives (CS_within) will be less likely than within-event transitions in FS narratives (FS_within) or across-event transitions in CS narratives (CS_across).

We predicted the probability of making correct forward recall transition using a logistic mixed-effects model, with the fixed effects of narrative type (CS vs. FS narrative), transition type (within- vs. across-event transition), and their interaction. For the narrative type predictor, the CS narrative condition was the reference condition. For the transition type predictor, the across-event transition condition was the reference condition. The model comparison and random effects selection process led us to retain the random intercept of subject and narrative as random effects. In Wilkinson notation, the model can be described as follows*: Accuracy of Transition Order (0/1) ~ Narrative Type + Transition Type + Narrative Type:Transition Type + (1 | Participant) + (1| Narrative)*.

Figure 4(D) shows the predicted probability for forward recall transition as a function of narrative type and transition type. The logistic mixed-effects model indicated a significant main effect of narrative type, $X^2(1) = 6.37$, $p = 0.01$, and a significant interaction between transition type and narrative type, $X^2(1) = 36.17$, $p < .001$. The main effect of transition type was not significant, $X^2(1) = 1.04$, $p = .31$. We further probed the interaction and tested the two hypothesized pairwise contrasts: We found that CS_within transitions had lower probability of being in the correct order than both CS_across pairs (B = -2.12, SE = 0.37, $p < .001$) and FS_within pairs (B = -3.09, SE = 0.56, $p < .001$).

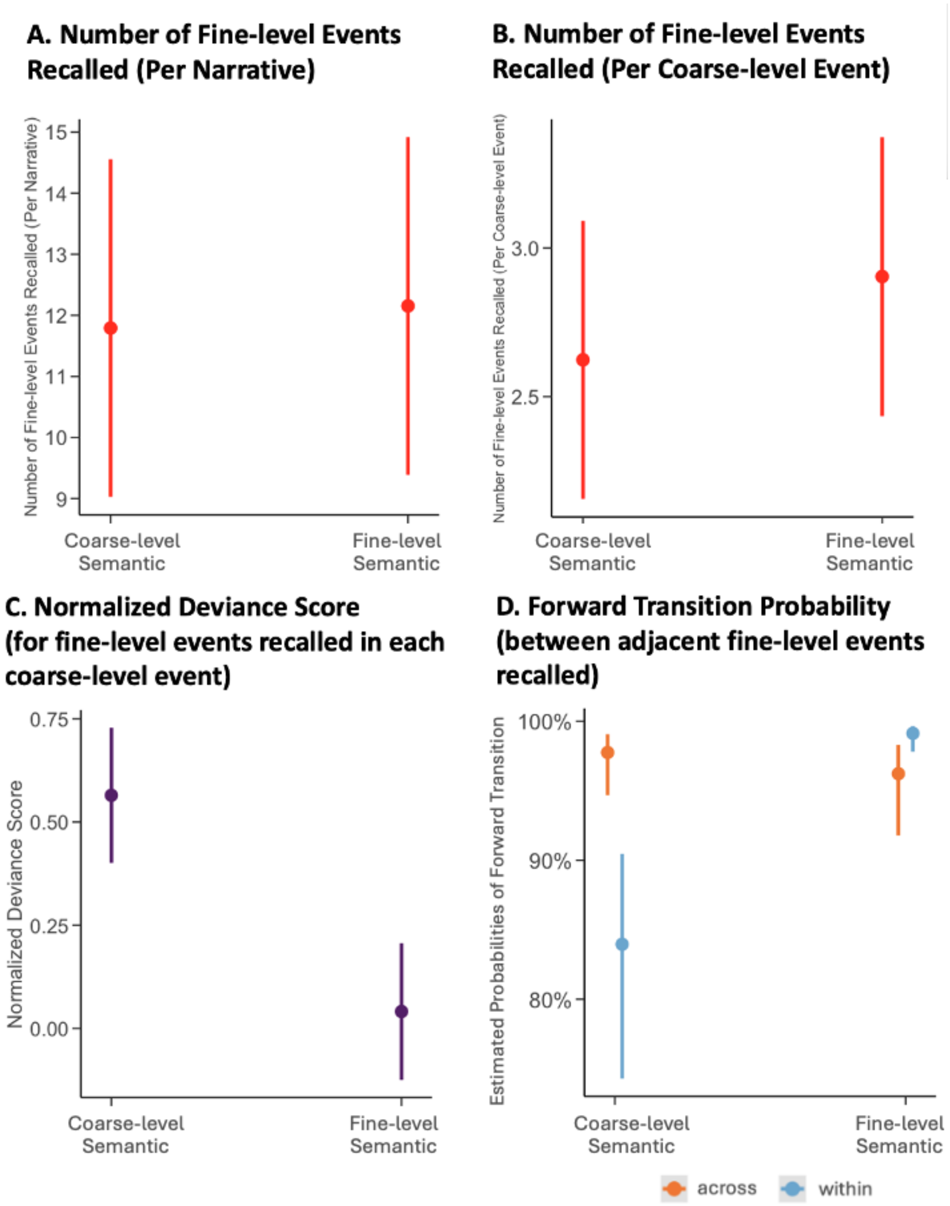


Figure 4. Experiment 3 results. (A) Estimated number of recalled fine-level events per narrative as a function of narrative type. (B) Estimated number of recalled fine-level events per coarse-level event as a function of narrative type. (C) Estimated normalized deviance score for fine-level events recalled in each coarse-level event. (D) Estimated forward transition probability as a function of narrative type and transition type. Error bars represent 95% confidence intervals.

4.3 Discussion

In Experiment 3, we examined how semantic order knowledge and hierarchical event structure influenced the temporal organization of events in long-term memory using a serial recall task. Our results suggested that fine-level semantic order knowledge exerted a strong influence in the order of fine-level events recalled within each coarse-level narrative: When there were semantic order constraints in the order of fine-level events, they were more likely to be recalled in order than when there were no semantic order constraints. This confirmed previous findings that event schema or scripts served as an important factor biasing the order of events recalled (Lichtenstein & Brewer, 1980; Bower et al., 1979; Brewer & Dupree, 1983).

In addition, by analyzing the order of adjacent recall transitions and looking at the influence of event boundary, we found evidence that narrative recall was organized hierarchically, with coarse-level event membership being an important grouping factor for the recall of fine-level events. First, we found that for adjacent fine-level events recalled in CS narratives, if they belonged to the same coarse-level event, they were less likely to be recalled in the correct order. But for adjacent fine-level events recalled in the FS narratives, if they belonged to the same coarse-level event, they were almost always recalled in the correct order. These two pieces of evidence offered additional support to the previous conclusion that fine-level semantic order constraint helped organize the order of recall, and correct order was not always successfully reconstructed when there was no semantic order knowledge facilitation. Second, we found that if adjacent fine-level events recalled in CS narratives belonged to different coarse-level events, they were almost always recalled in the correct order. To some extent, this suggests that participants formed different event models for each coarse-level event and organized them into the correct order in the long-term memory using the coarse-level semantic constraints that we provided to CS narratives as a manipulation. When recalling the narrative, they would follow the hierarchical organization of event models, recalling all fine-level events they could remember from the first coarse-level event, and then move on to the second coarse-level event and recall all fine-level events they could remember there, and continue this process until they scanned all the coarse-level events they could remember. This would explain why we almost never observe CS_across transitions in the wrong order, and the proposed process corresponds to one of the principles proposed in the Event Horizon Model: Causal connectivity is the dominant factor organizing the relationship among event models the long-term memory (Radvansky, 2012; Radvansky & Zacks, 2017). In addition, we observed a surprising result that FS_across transitions in serial recall also had a very high (more than 95%) probability of being in the correct direction, almost as high as CS_across and FS_within transitions. At first glance, this might seem to contradict the results in Experiment 1 and 2 that recency judgment for FS_across pairs consistently had impaired performance comparing to CS_across and FS_within pairs with semantic order knowledge facilitation. However, we need to consider the fact that the recall transition analysis in Experiment 3 was conducted based on the events that were recalled by the participants, which did not include everything in the narrative. In fact, participants recalled about half of the fine-level events (mean = 12.01) in each narrative narrative. One possibility is that for the proportion of the coarse-level events that they remembered, participants had very accurate episodic memory of their relative order. These event models were organized in the long-term memory by strong episodic associations that formed during encoding, which led to an accurate order reconstruction during retrieval. However, some FS_across event pairs we tested in recency judgment tasks might come from the coarse-level events they did not remember, which led to inaccurate performance for these pairs. A second possibility is that participants had very accurate episodic memory for almost all the coarse-level events, which led to the accurate FS_across transitions that we observed in Experiment 3. However, it is possible that for some FS_across sentences that we tested in recency judgment tasks, participants had inaccurate source memory and linked them to the coarse-level events they did not belong to, which caused impaired performance. A third possibility is that participants did not consistently use order information among coarse-level events to inform their decision for fine-level recency judgments in Experiment 1 and 2, despite the fact that it could have supported accurate performance. Follow-up experiments need to be conducted to distinguish between these possibilities.

In the current experiment, we did not observe the hypothesized effect that more fine-level events would be recalled per narrative (or per coarse-level event) in FS narratives compared to in CS narratives. Since fine-level events in FS narratives were mostly causally connected within each coarse-level event, participants were likely to jump over some events in the causal chain, while still maintaining the correct forward transition order. For example, when participants were reconstructing the *"visit the snack cart"* coarse-level event of the *"visiting the zoo"* FS narrative, after recalling the protagonist "*waited in the line*," they might skip the "*told the owner what he wanted*" event to report "*paid with cash and received the food*." Omissions like this might mask the difference in the number of units recalled across the two conditions.

Chapter 5: General Discussion

In this set of experiments, we developed a novel narrative reading paradigm to examine how semantic order knowledge and hierarchical event structure can be leveraged to facilitate the reconstruction of temporal order relationship among events. To summarize, we first demonstrated that coarse-level event changes in our narratives were reliably perceived as event boundaries. This was characterized by increased reading time for boundary sentences during encoding (Experiment 1) and inflated temporal distance rating for fine-level events pairs that spanned across event boundary during retrieval (Experiment 1 and 2), which replicated findings in the previous literature (Radvansky et al., 2001; Zwaan & Radvansky, 1998; DuBrow & Davachi, 2013; Ezzyat & Davachi, 2014). Furthermore, we confirmed our critical hypothesis that semantic order knowledge on both coarse and fine levels could be applied to facilitate recency judgment, and this effect overrode the influence of event boundary alone (Experiment 1 and 2). Although many previous picture-list learning paradigms found a consistent effect that event boundary impaired recency judgment for items that spanned across a contextual boundary (DuBrow & Davachi, 2013; Heusser et al., 2018; Pu et al., 2022), our results showed that at least in more naturalistic events, such as narratives about everyday activities, the effect of event boundary on recency judgment depended on people's ability to use information other than episodic memory, which included their semantic knowledge about stereotypical event order. When there were semantic order constraints among fine-level events, recency judgment of fine-level events within the same coarse-level event was improved, suggesting that people could use direct fine-level order knowledge to facilitate temporal order reconstruction on the same level. However, the mechanism was slightly different in another scenario: When there were semantic order constraints among coarse-level events, recency judgment of fine-level events coming from different coarse-level events were improved. This cross-level semantic facilitation suggested that participants might form a hierarchical representation of the narrative in their long-term memory, which enabled the order among coarse-level events to inform the reconstruction of the order among fine-level events. This was further supported by results from the recency judgment confidence measure: When there was no semantic order knowledge facilitation, having coarse-level event information as an extra source of information increased confidence in recency judgment, after controlling for accuracy (Experiment 1 and 2). By analyzing participants' serial recall of the narrative, we found further evidence that participants frequently chunked their recall of fine-level events based on their coarse-level event membership (Experiment 3). In addition, we also discovered that semantic order constraints served as an important factor in organizing the order of recall on both the coarse-level and fine-level.

Our results were largely consistent with the principles outlined by Event Segmentation Theory (Zacks et al., 2007) and Event Horizon Model (Radvansky, 2012; Radvansky & Zacks, 2017) in how people create structured representations of events during perception and store them into long-term memory. These theoretical frameworks argued that people segment an ongoing stream of activity into distinct event models and transform them into "episodes" in long-term memory. During this process, semantic knowledge not only plays an important role in combining with incoming sensory information to construct the working event model, but also serves as an important factor in organizing the relationship among event models in long-term memory (Radvansky & Zacks, 2017; Zacks, 2020). In addition, it is worth noting that one role event segmentation plays in shaping long-term memory is that the event structure formed during perception can serve as an effective chunking mechanism. By grouping fine-level events into larger chunks of coarse-level events, it helped with creating a hierarchical and relational representation of events in long-term memory. This representation is likely to include a mechanism to represent the order of events both on the coarse-level and fine-level, and a way to represent the membership of fine-level events within certain coarse-level events. In this way, when asked to judge the recency between two fine-level events that come from two different coarse-level events, instead of only relying on direct temporal relationship between the two fine-level events, an additional way is to consult the temporal relationship between the coarse-level events they each belong to. Since the chunking mechanism of event segmentation determined that there will be less coarse-level events than fine-level events, it is adaptive to have access to a mechanism that requires encoding less temporal relationship among individual units.

We plan to look for behavioral evidence for the mechanism outlined above in a follow-up experiment, in which we aimed to figure out different sources of information used in the recency judgment of fine-level pairs that each belonged to a different coarse-level event. First, we plan to conduct an additional source memory test to see if participants can link each fine-level sentence to the coarse-level event it belongs to. Second, we plan to test coarse-level recency judgment by asking participants to order of the two coarse-level events that the fine-level sentences belong to. If the hierarchical mechanism proposed above is correct, we should see that the accuracy of fine-level recency judgment being strongly predicted by both source memory accuracy and coarse-level recency accuracy.

It is worth noting that the hierarchical mechanism that we outlined here has some fundamental differences from previous computational models that were proposed for temporal order memory and recency judgment (Horner et al., 2016; Pu et al., 2022). Those models were largely based on picture-list learning paradigms that orthogonalized the relationship between “items” and “contexts,” and the only role that event boundaries play in these paradigms is to alter the stability of the encoding context. In their implementation, both of the two models were built on a class of temporal context models (Estes, 1950; Howard & Kahana, 2002), which associate different items to be encoded with a context signal that gradually drifts over time. Both models assumed a sharp change in the contextual representation at event boundary, but one implemented the change with a faster random drifting rate in the context signal (Horner et al., 2016), and another implemented the change with a reinstatement of the pre-experimental context (Pu et al., 2022). However, since “context” was implemented as a continuously drifting signal in these models, there was no explicit representation of each context (e.g., color background, task type, etc.) that was present in the behavioral paradigms, which made them hard to accommodate behavioral results from our experiments using hierarchical narrative stimuli. We argue that in order to build computational models that account for temporal order memory phenomena in naturalistic context, we need to incorporate mechanisms for hierarchical reasoning and the reuse of schematic structures into the model architecture.

Another implication for the paradigm we developed here is to use it for examining how aging influences people’s ability to encode and reconstruct temporal relationships among everyday events. Using highly controlled stimuli, some previous studies of older adults’ temporal order memory have distinguished their intact ability to utilize pre-experimental semantic association from their general deficit to encode and use new episodic associations online. A study looking at the serial recall of word lists found that older adults relied more heavily than younger adults on semantic organization among words to structure recall order, and made fewer in-order recall transitions based on newly encoded temporal context (Golomb et al., 2008). Similarly, a previous study using short sequences of pictures found that older adults’ performance in identifying out-of-sequence items was comparable to younger adults when tested on familiar and predictable sequences (e.g. the letters A-F), but was greatly impaired for newly encountered sequences (e.g. distinct fractal images) (Allen et al., 2015). These findings suggested that specific types of sequential processing might be differentially affected by aging, depending on whether the task allows older adults to leverage their pre-existing knowledge structure. In the context of our paradigm, it is natural to hypothesize that when doing recency judgment tasks, the existence of semantic order knowledge will provide a greater facilitation to older adults' temporal order judgment and recall memory, comparing to younger adults. By testing this hypothesis using stimuli that closely approximate real-life scenario, we might be able to develop more targeted interventions that exploit existing knowledge structure to improve older adults' temporal memory in everyday tasks.

One potential limitation of the current study is that the stimuli we used contain pre-determined hierarchical and semantic structure created by the experimenter, such that a given narrative can only contain either coarse-level or fine-level semantic order constraints. In real life scenario, the event dynamic we attempt to reconstruct from autobiographical memory is likely to be more nuanced, with a hierarchical structure of more than two levels and a mix and match of semantic order constraints on different levels. This concern needs to be addressed in future studies, in which real-world stimuli including movies or novels can be used to test the conclusions made by the current study.

In conclusion, Experiment 1 and 2 suggest that semantic order knowledge and hierarchical event structure can be leveraged to reconstruct the order of naturalistic events. Experiment 3 suggests that coarse-level event membership serves as chucking mechanism to order fine-level events in serial recall. Together, the findings from all three experiments highlight that the reconstruction of the temporal order among events can depend on many sources of information coming from semantic and episodic memory, across different time scales.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist*, *36*(7), 715–729. https://doi.org/10.1037/0003-066X.36.7.715

Allen, T. A., Morris, A. M., Stark, S. M., Fortin, N. J., & Stark, C. E. L. (2015). Memory for sequences of events impaired in typical aging. *Learning & Memory*, *22*(3), 138–148. https://doi.org/10.1101/lm.036301.114

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology* (pp. xix, 317). Cambridge University Press.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious Mixed Models* (arXiv:1506.04967). arXiv. https://doi.org/10.48550/arXiv.1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*(2), 177–220. https://doi.org/10.1016/0010-0285(79)90009-4

Brewer, W. F., & Dupree, D. A. (1983). Use of plan schemata in the recall and recognition of goal-directed actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 117–129. https://doi.org/10.1037/0278-7393.9.1.117

Clewett, D., Gasser, C., & Davachi, L. (2020). Pupil-linked arousal signals track the temporal organization of events in memory. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17851-9

Davis, E. E., & Campbell, K. L. (2023). Event boundaries structure the contents of long-term memory in younger and older adults. *Memory*, *31*(1), 47–60. https://doi.org/10.1080/09658211.2022.2122998

Diamond, N. B., & Levine, B. (2020). Linking Detail to Temporal Structure in Naturalistic-Event Recall. *Psychological Science*, *31*(12), 1557–1572. https://doi.org/10.1177/0956797620958651

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, *142*(4), 1277–1286. https://doi.org/10.1037/a0034024

DuBrow, S., & Davachi, L. (2014). Temporal Memory Is Shaped by Encoding Stability and Intervening Item Reactivation. *Journal of Neuroscience*, *34*(42), 13998–14005. https://doi.org/10.1523/JNEUROSCI.2535-14.2014

DuBrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017). Does mental context drift or shift? *Current Opinion in Behavioral Sciences*, *17*, 141–146. https://doi.org/10.1016/j.cobeha.2017.08.003

Estes, W. K. (1950). Toward a statistical theory of learning: Psychological Review. *Psychological Review*, *57*(2), 94–107. https://doi.org/10.1037/h0058559

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, *22*(2), 243–252. https://doi.org/10.1177/0956797610393742

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin*, *113*(1), 44–66. https://doi.org/10.1037/0033-2909.113.1.44

Gernsbacher, M. A. (1991). Cognitive Processes and Mechanisms in Language Comprehension: The Structure Building Framework. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 27, pp. 217–263). Academic Press. https://doi.org/10.1016/S0079-7421(08)60125-5

Golomb, J. D., Peelle, J. E., Addis, K. M., Kahana, M. J., & Wingfield, A. (2008). Effects of adult aging on utilization of temporal and semantic associations during free and serial recall. *Memory & Cognition*, *36*(5), 947–956. https://doi.org/10.3758/MC.36.5.947

Gurguryan, L., Dutemple, E., & Sheldon, S. (2021). Conceptual similarity alters the impact of context shifts on temporal memory. *Memory*, *29*(1), 11–20. https://doi.org/10.1080/09658211.2020.1841240

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, *34*(6), 1221–1235. https://doi.org/10.3758/BF03193267

Heusser, A. C., Ezzyat, Y., Shiff, I., & Davachi, L. (2018). Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1075–1090. https://doi.org/10.1037/xlm0000503

Hintzman, D. L. (2005). Memory strength and recency judgments. *Psychonomic Bulletin & Review*, *12*(5), 858–864. https://doi.org/10.3758/BF03196777

Horner, A. J., Bisby, J. A., Wang, A., Bogus, K., & Burgess, N. (2016). The role of spatial boundaries in shaping long-term event representations. *Cognition*, *154*, 151–164. https://doi.org/10.1016/j.cognition.2016.05.013

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31965-2

Lewandowsky, S., & Murdock, B. B. Jr. (1989). Memory for serial order. *Psychological Review*, *96*(1), 25–57. https://doi.org/10.1037/0033-295X.96.1.25

Lichtenstein, E. H., & Brewer, W. F. (1980). Memory for goal-directed events. *Cognitive Psychology*, *12*(3), 412–445. https://doi.org/10.1016/0010-0285(80)90015-8

McClay, M., Sachs, M., & Clewett, D. (2022). *Dynamic music-induced emotions shape the episodic structure of memory* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8hpwy

McNerney, M. W., Goodwin, K. A., & Radvansky, G. A. (2011). A Novel Study: A Situation Model Analysis of Reading Times. *Discourse Processes*, *48*(7), 453–474. https://doi.org/10.1080/0163853X.2011.582348

McRae, K., Brown, K. S., & Elman, J. L. (2021). Prediction-Based Learning and Processing of Event Knowledge. *Topics in Cognitive Science*, *13*(1), 206–223. https://doi.org/10.1111/tops.12482

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review*, *90*(4), 316–338. https://doi.org/10.1037/0033-295X.90.4.316

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Pu, Y., Kong, X.-Z., Ranganath, C., & Melloni, L. (2022). Event boundaries shape temporal organization of memory by resetting temporal context. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-28216-9

Radvansky, G. A. (2012). Across the Event Horizon. *Current Directions in Psychological Science*, *21*(4), 269–272. https://doi.org/10.1177/0963721412451274

Radvansky, G. A., & Copeland, D. E. (2010). Reading times and the detection of event shift processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 210–216. https://doi.org/10.1037/a0017258

Radvansky, G. A., & Zacks, J. M. (2014). *Event Cognition*. Oxford University Press.

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140. https://doi.org/10.1016/j.cobeha.2017.08.006

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and aging. *Psychology and Aging*, *16*(1), 145–160. https://doi.org/10.1037/0882-7974.16.1.145

Rinck, M., & Weber, U. (2003). Who when where: An experimental test of the event-indexing model. *Memory & Cognition*, *31*(8), 1284–1292. https://doi.org/10.3758/BF03195811

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, *203*, 104269. https://doi.org/10.1016/j.cognition.2020.104269

Rubin, D. C., & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review*, *122*(1), 1–23. https://doi.org/10.1037/a0037907

Sasmita, K., & Swallow, K. M. (2022). Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01832-5

Sols, I., DuBrow, S., Davachi, L., & Fuentemilla, L. (2017). Event Boundaries Trigger Rapid Memory Reinstatement of the Prior Events to Promote Their Representation in Long-Term Memory. *Current Biology*, *27*(22), 3499-3504.e4. https://doi.org/10.1016/j.cub.2017.09.057

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381–391. https://doi.org/10.1016/S0022-5371(66)80048-8

Van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension.

Wang, Y. C., & Egner, T. (2022). Switching task sets creates event boundaries in memory. *Cognition*, *221*, 104992. https://doi.org/10.1016/j.cognition.2021.104992

Wen, T., & Egner, T. (2022). Retrieval context determines whether event boundaries impair or enhance temporal order memory. *Cognition*, *225*, 105145. https://doi.org/10.1016/j.cognition.2022.105145

Zacks, J. M. (2020). Event Perception and Memory. *Annual Review of Psychology*, *71*(1), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101

Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the Alzheimer type. *Psychology and Aging*, *21*(3), 466–482. http://dx.doi.org.proxy.library.vanderbilt.edu/10.1037/0882-7974.21.3.466

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, *138*(2), 307–327. https://doi.org/10.1037/a0015305

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162

Appendix A: Supplementary Methods

A.1 Norming Study

To create Coarse-level Semantic (CS) and Fine-level Semantic (FS) narratives, we conducted a norming study using a separate sample to confirm that the event sets with semantic order constraints were agreed by people to happen only in one specific order, and that the event sets without semantic order constraints were agreed by people to potentially happen in any order.

The stimuli used in the norming study were sixty sets of event labels describing familiar everyday activities, with five event labels in each set. Ten sets were coarse-level event labels sets that corresponded to the five coarse-level events in each of the ten narratives. The remaining fifty sets were fine-level event label sets that corresponded to each of the five sets of fine-level event sentences in each narrative.

Our sample included 36 participants recruited from Prolific (https://www.prolific.com/). During the online study, participants were shown different sets of event labels describing familiar everyday activities. They were asked to answer two questions for each set of event labels: (1) Rating: To rate the degree to which these activities should occur in a certain temporal order (scale = 1-7, 1 = "could happen in any order," 7 = "could happen in only one order"), and (2) Ranking: To sort the five activities into what they believe to be the most likely order, regardless of how they answered the rating question. The order of event label sets was randomized for each participant, and the layout order of event labels within each set was randomized for each trial.

To quantify how much the event order ranking provided by the participants deviated from the order they were presented in the narrative, we calculated order deviance scores based on participants' answers to the ranking questions. For example, if the order of one set of event labels was [1, 2, 3, 4, 5] in the narrative, and the ranking answer a participant gave was [3, 1, 2, 4, 5], deviance score = |1-3| + |3+1-1| + |1+1-2| + |2+1-3| + |4+1-5| = 5. The range of the deviance score is [0, 15], with 0 meaning that participants' ranking was exactly the same as the order in the narrative, 15 meaning that participants' ranking was completely the opposite comparing to the order in the narrative.

Therefore, for the rating question, we hypothesized that both coarse-level event label sets from CS narratives and fine-level event label sets from FS narratives should be mostly rated as "could happen in only one order," with values close to 7. In contrast, both fine-level event label sets from CS narratives and coarse-level event label sets from FS narratives should be mostly rated as "could happen in only one order," with values close to 1.

For the ranking question, we hypothesized that both coarse-level event label sets from CS narratives and fine-level event label sets from FS narratives should have low order deviation scores, with values close to 0. In contrast, both fine-level event label sets from CS narratives and coarse-level event label sets from FS narratives should have high order deviation scores, with values farther away from 0.

In Table 2, we report the mean rating scores for both coarse-level and fine-level event labels sets in each narrative, with their corresponding 95% confidence interval. In Table 3, we report the mean ranking deviance score for both coarse-level and fine-level event labels sets in each narrative, with their corresponding 95% confidence interval. Based on the results, we can conclude that our manipulation worked as intended for developing narrative stimuli: According to their semantic knowledge, people agreed that coarse-level events in CS narratives and fine-level events within each coarse-level event in FS narratives should follow a specific order, and that fine-level events within each coarse-level event in CS narratives and coarse-level event in FS narratives did not need to follow a specific order.

	Coarse-level event label set (*1)	Fine-level event label sets (*5)
CS1_Aunt	6.08 [5.65, 6.51]	2.30 [2.05, 2.55]
CS2_Swimming	6.58 [6.18, 6.98]	2.69 [2.41, 2.97]
CS3_Morning	6.25 [5.90, 6.60]	2.14 [1.92, 2.36]
CS4_Cafeteria	6.33 [5.88, 6.79]	2.24 [1.97, 2.52]
CS5_Examination	6.08 [5.64, 6.52]	1.98 [1.74, 2.22]
FS1_Shopping	1.42 [1.25, 1.58]	6.31 [5.93, 6.69]
FS2_Cleaning	1.53 [1.33, 1.72]	5.77 [5.20, 6.35]
FS3_Zoo	1.50 [1.31, 1.69]	6.51 [6.17, 6.84]
FS4_Campus	1.36 [1.24, 1.48]	6.32 [5.90, 6.74]
FS5_Farm	1.47 [1.27, 1.67]	6.44 [6.04, 6.85]

Table 2. Mean rating scores for both coarse-level and fine-level event labels sets in each narrative, with their corresponding 95% confidence interval. Note that for each narrative, there was only one coarse-level event label set and five fine-level event sets. For the name of each narrative, “CS” meant that it was a CS narrative, and “FS” meant that it was a FS narrative.

	Coarse-level event label set (*1)	Fine-level event label sets (*5)
CS1_Aunt	0.81 [0.03, 1.58]	10.40 [9.98, 10.82]
CS2_Swimming	1.06 [0.40, 1.71]	9.07 [8.65, 9.50]
CS3_Morning	1.36 [0.42, 2.30]	10.59 [10.18, 11.01]
CS4_Cafeteria	0.19 [0, 0.47]	9.81 [9.32, 10.30]
CS5_Examination	2.22 [1.08, 3.36]	9.81 [9.37, 10.24]
FS1_Shopping	10.78 [10.37, 11.18]	1.14 [0.37, 1.91]
FS2_Cleaning	9.42 [8.94, 9.89]	2.12 [1.05, 3.18]
FS3_Zoo	10.22 [9.84, 10.60]	1.27 [0.48, 2.07]
FS4_Campus	10.19 [9.82, 10.57]	1.06 [0.19, 1.93]
FS5_Farm	9.78 [9.21, 10.35]	0.71 [0, 1.44]

Table 3. Mean ranking order deviance scores for both coarse-level and fine-level event labels sets in each narrative, with their corresponding 95% confidence interval. Note that for each narrative, there was only one coarse-level event label set and five fine-level event sets. For the name of each narrative, "CS" meant that it was a CS narrative, and "FS" meant that it was a FS narrative.

A.2 Semantic Similarity Analysis

For the recency judgment task that was used in Experiment 1 and 2, we acknowledged that the accuracy of recency judgment could potentially be affected by a confounding variable, which was the semantic similarity between the two sentences in each event pair. Therefore, we used Universal Sentence Encoder (USE) to convert each event sentence being tested into a sentence embedding and calculated the inner product of the two sentence embeddings in each event pair to quantify their semantic similarity. We then compared the semantic similarity between the two sentences in each event pair across four different types of pairs.

An one-way between-groups ANOVA revealed that there was not a significant effect of pair type on semantic similarity, *F*(3, 76) = 1.084, *p* = 0.36. We then calculated Bayes factors for the four pairwise contrasts that was hypothesized and tested in the main analysis: We found anecdotal evidence that there was not a significant difference in semantic similarity between FS_within and CS_within conditions ($BF_{10} = 0.31$), between FS_within and FS_across conditions ($BF_{10} = 0.33$), between CS_across and FS_across conditions ($BF_{10} = 0.53$), and between CS_across and CS_within conditions ($BF_{10} = 0.85$). Based on this result, we concluded that semantic similarity did not differ systematically across event pairs in different conditions and was not likely to have caused the recency judgment accuracy differences we observed.

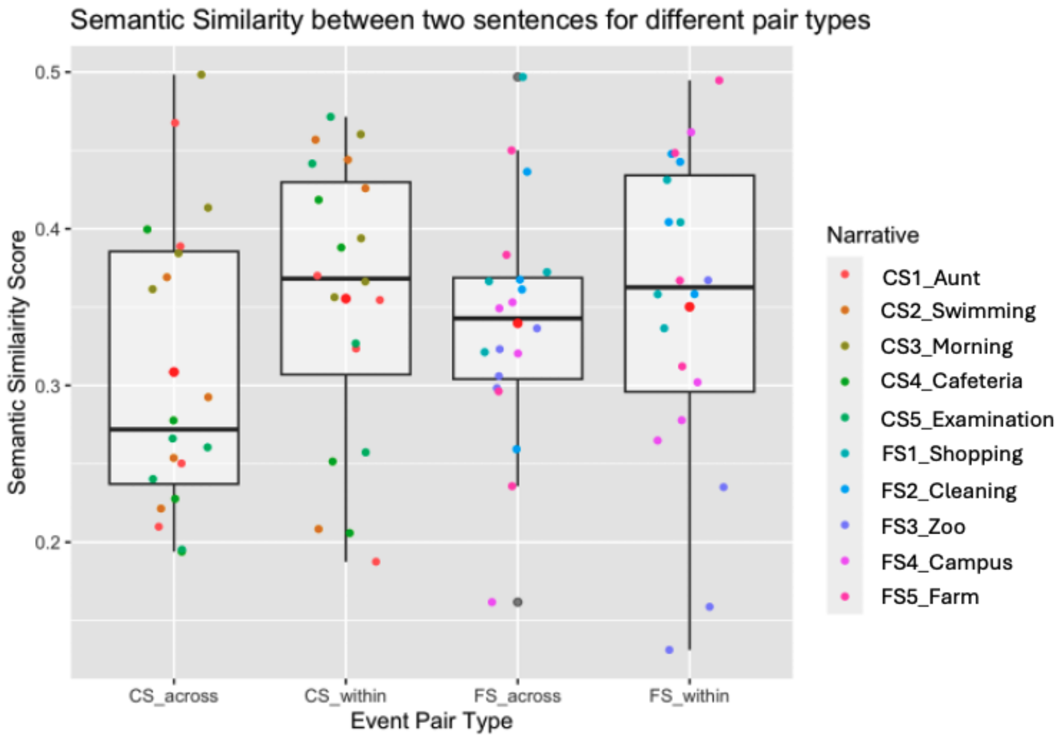


Figure 5. Semantic similarity between two sentences in each event pair, across four different event pair types. Each dot on the graph represents a different fine-level event sentence pair that was tested in Experiment 1 and 2. The red dot in the center of each bar represents the mean semantic similarity score within each event pair type.