PNAS RESEARCH ARTICLE | NEUROSCIENCE

Emergence of a temporal processing gradient from naturalistic inputs and network connectivity

Claire H. C. Chang[a,b,1], Samuel A. Nastase[a], Uri Hasson[a], and Peter Ford Dominey[c,1]

Edited by Marcus Raichle, Washington University in St. Louis School of Medicine, St. Louis, MO; received October 1, 2024; accepted May 15, 2025

Natural language unfolds over multiple nested timescales: Words form sentences, sentences form paragraphs, and paragraphs build into full narratives. Correspondingly, the brain exhibits a hierarchy of processing timescales, spanning from lower- to higher-order regions. During narrative comprehension, neural activation patterns have been shown to propagate along this cortical hierarchy with increasing temporal delays (lags). To investigate the mechanisms underlying this lag gradient, we systematically manipulate the structure of a recurrent reservoir network. In the biologically inspired "Limited-Canal" configuration, word embeddings are received by a limited set of sensory neurons and transmitted through a series of local connections to the distal end of the network. This configuration endows the network with an intrinsic lag gradient, inducing a cascade of activity as information propagates along the network. We found that, similar to the human brain, this intrinsic lag gradient is enhanced by naturalistic narratives. The interaction between naturalistic input and network structure becomes evident when manipulating local connectivity through the "canal width" parameter, which determines how closely the Limited-Canal model mirrors the human brain's sensitivity to narrative structure. In addition, we found that processing cost, as a computational proxy for the BOLD signal, increases more slowly in later neurons, which can account for the emergence of the lag gradient. Our results demonstrate that narrative-driven neural dynamics can emerge from macroscale anatomical topology alone without task-specific training. These fundamental topological properties of the human cortex may have evolved to effectively process the hierarchical structures ubiquitous in the natural environment.

temporal processing hierarchy | naturalistic narrative | fMRI | recurrent network | reservoir

Humans are remarkably sensitive to the complex, nested structures of events in the external world. A growing line of work using naturalistic narrative stimuli (e.g., spoken stories) has revealed a cortical processing hierarchy thought to encode these structures (1–8). The structure of natural language is temporally nested over multiple timescales, where phonemes build words, words build sentences, sentences build paragraphs, and paragraphs combine to create narratives. Along the temporal processing hierarchy, areas responsible for processing speech in the superior temporal gyrus (STG) appear to rapidly integrate information over hundreds of milliseconds, which corresponds to the integration of phonemes into words. Adjacent areas along the STG seem to integrate information over seconds, which corresponds to the integration of words into phrases and sentences (9). Finally, areas in the default mode network (DMN) at the top of the processing hierarchy appear to integrate information over tens of seconds, which corresponds to the integration of information across paragraphs (10–12).

The temporal processing hierarchy was found using different linguistic stimuli and was observed while using spoken or written narratives (12–16). Chien and Honey (4) found a gradual flow of information across the cortical hierarchy. In a clever design, they measured activity alignment to a given paragraph in two groups of subjects who listened to two different preceding stories. The time it took for their neural responses to converge served as an estimation of the integration time constant. The results indicate that sensory cortices with a short temporal integration window aligned most quickly across subjects to shared paragraphs, followed by mid-level regions with an intermediate temporal integration window. At the top of the temporal integration hierarchy, it took more than 10 s to align to information conveyed along the paragraph. This corroborates work by Baldassano et al. (17) demonstrating a hierarchy of cortical event representations. They observed that narrative stimuli evoke more numerous and shorter neural events in sensory cortices, while higher-level areas, such as the posterior medial cortex, exhibited fewer and longer events (see also refs. 18 and 19).

Infant studies provide an opportunity to investigate the development of the hierarchically organized processing timescales. It has been shown that infants exhibit reliable

Significance

We manipulate the structure of an artificial neural network to isolate the influence of anatomical architecture on the emergence of the topological gradient of activation pattern propagation observed in the human brain. Using networks that process words sequentially and encode their relationships through recurrent activities, we found that the propagation of external information from sensory neurons through local connections yields an intrinsic lag gradient. Notably, even without task-specific training, this biologically inspired architecture exhibits an enhanced lag gradient when processing naturalistic narratives, similar to human brain dynamics. These findings suggest that evolutionary pressure might have been the drive underlying the development of such an architecture.

Author affiliations: [a]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540; [b]The Graduate Institute of Mind, Brain, and Consciousness, College of Humanities and Social Sciences, Taipei Medical University, New Taipei City 235, Taiwan; and [c]Université Bourgogne Europe, Inserm, Cognition, Action et Plasticité Sensorimotrice, Unité Mixte de Recherche 1093, Dijon 21000, France

Author contributions: C.H.C.C. and P.F.D. designed research; C.H.C.C. and P.F.D. performed research; C.H.C.C. and P.F.D. analyzed data; and C.H.C.C., S.A.N., U.H., and P.F.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: claire.hc.chang@gmail.com or peter.dominey@inserm.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2420105122/-/DCSupplemental.

Published July 9, 2025.

processing timescales during both rest (20) and cartoon watching (21), although these timescales do not exhibit the hierarchical organization from sensory to transmodal regions. However, disentangling the contributions of cortical anatomical immaturity and limited prior experience remains a challenge when comparing infants to adults.

As a complementary approach, artificial neural networks, though lacking certain biological features of the brain, offer a means to systematically compare neural computation with varying anatomical and functional architectures, features of large-scale network organization that cannot be directly manipulated in humans. Studies comparing the output of artificial networks with human behavior and neural responses have gained increasing attention (22–27).

Artificial neural networks encode linguistic inputs (e.g., words) as distributed activity vectors across units, comprising a high-dimensional embedding space where geometric relations among vectors capture linguistic relationships among words (e.g., ref. 28). Embeddings extracted from higher network layers seem to code semantic and syntactic relationships across longer temporal scales (29, 30). Among the various architectures of artificial neural networks, the transformer architecture has come to dominate natural language processing tasks in recent years (e.g., refs. 31–33). Transformers rely on repeated layers of a particular circuit motif (the "self-attention head") that integrates information across words in a fixed context window to adjust the meaning of each word in context. Recent work has demonstrated that transformer-based models learn internal representations that can predict brain activity measured during naturalistic narrative comprehension with remarkable accuracy (26, 34–38). However, while humans must process linguistic inputs sequentially (39), transformers simultaneously process an entire input sequence—up to the length of the context window [for example, 1,024 tokens in GPT-2 (32), and tens of thousands of tokens in more recent models, e.g., GPT-4].

On the other hand, recurrent neural networks process one word at a time (40, 41). Similarly to humans, recurrent networks can encode temporal relationships among words, which makes them particularly relevant to human biological language processing (38, 42). Sequential input in recurrent networks produces waves of signals that overlap and evolve in time, which encode the spatial and temporal features of the input (43). Reservoir computing (44) employs recurrent connections (projections from neurons in one layer back onto other neurons in the same layer) to generate rich dynamics that have powerful computational properties (45). Local recurrent connectivity is a principal feature of the primate cortex (46). In the reservoir network, the continuous contextual history of past words is maintained in the flow of their trace through the recurrent connections. Task-specific training only modifies the output weights, not the internal recurrent connections. In other words, training affects the extraction of contextual memory but not its structure and maintenance in the reservoir model. The resulting capabilities for temporal integration and high-dimensional representation allow reservoir models to capture relevant properties of cortical dynamics at the level of a single unit and neuronal population activity (47). Reservoir computing has recently been used to examine cognitive function and brain dynamics with architectures constrained by the human connectome (48–50). While artificial neural networks do not capture many of the detailed biological features of the brain, reservoir networks capture the recurrent dynamics, making them particularly suitable for modeling the neural dynamics within the cortical processing hierarchy (relative to, e.g., transformers or other feedforward artificial neural network models).

Indeed, given the sequential nature of human narrative processing, such recurrent networks with inherent temporal dynamics may provide novel insights into the brain dynamics underlying narrative comprehension. In a first effort in this direction (51), we used a recurrent reservoir model to reproduce findings of both the increasing narrative time constants (4) and the hierarchy of event representations (17) in silico. In a classical reservoir model, as reported in Chien and Honey (4), we observed a broad distribution of integration time constants, a measure of how long a neuron bin retains the influence of past input or how slowly it "forgets," and, as reported by Baldassano and colleagues (17), neural populations with longer time constants preferred longer narrative events. These findings suggest that these two processes derive from a common temporal integration mechanism inherent in reservoir computing (51). However, unlike the human brain, the classical reservoir structure lacks hierarchical topology, as all units receive sensory inputs and are connected to all other units. Neurons with different temporal integration constants are scattered across the network [Dominey et al. (52); reproduced in *SI Appendix*, Figs. S1 and S2].

Inspired by previous work indicating that biological neural networks are optimized for hierarchically structured naturalistic inputs (10, 52, 53), in this study, we test the hypothesis that reservoir networks may capture the key architectural properties of large-scale cortical organization and reproduce human brain dynamics evoked by naturalistic narratives. We targeted a topographic gradient of lagged activity across brain networks observed during narrative comprehension (54) (Fig. 1). Namely, similar neural response fluctuations appear across distinct functional networks with systematic temporal delays. To quantify stimulus-driven, interregional temporal lags, we first computed intersubject functional connectivity using a leave-one-subject-out approach. We then identified the lag that maximized functional connectivity between network pairs. This lag gradient follows the cortical processing hierarchy (10, 12), where higher-order areas with longer temporal integration windows have relatively longer activity delays than areas lower in the hierarchy with shorter temporal integration windows. In prior work (54), We proposed that this gradient reflects the recursive construction of increasingly larger linguistic or narrative events, where the output at one level (e.g., words) becomes the input for the next (e.g., phrases). Supporting this interpretation, this lag gradient is diminished when the temporal structure of a narrative is disrupted by temporally scrambling the narrative stimulus.

Our prior simulations demonstrated that nested linguistic structures are a crucial factor for the emergence of the lag gradient (54). We explicitly simulated the hierarchically structured linguistic/narrative event boundaries in naturalistic narratives and their temporal integration functions. Simulated neural activity in one processing level is passed to the next, with each level integrating information over longer timescales and resetting at event boundaries. This simple model was sufficient to produce the observed lag gradient. In other words, the lag gradient arises from the gradual integration and transmission of linguistic units (words, sentences, paragraphs) along the processing hierarchy.

In the current research, we examine whether the lag gradient can emerge from intrinsic structure in the network and the input, without explicitly specified narrative boundaries and response functions. To begin quantifying the semantic structure of the narrative texts, we utilized the Wikipedia2Vec model, pretrained on the three billion words 2018 Wikipedia corpus, to generate word embeddings from the narratives. These embeddings encode the semantic meaning of words such that semantically related words will have similar embeddings (corresponding to nearby locations

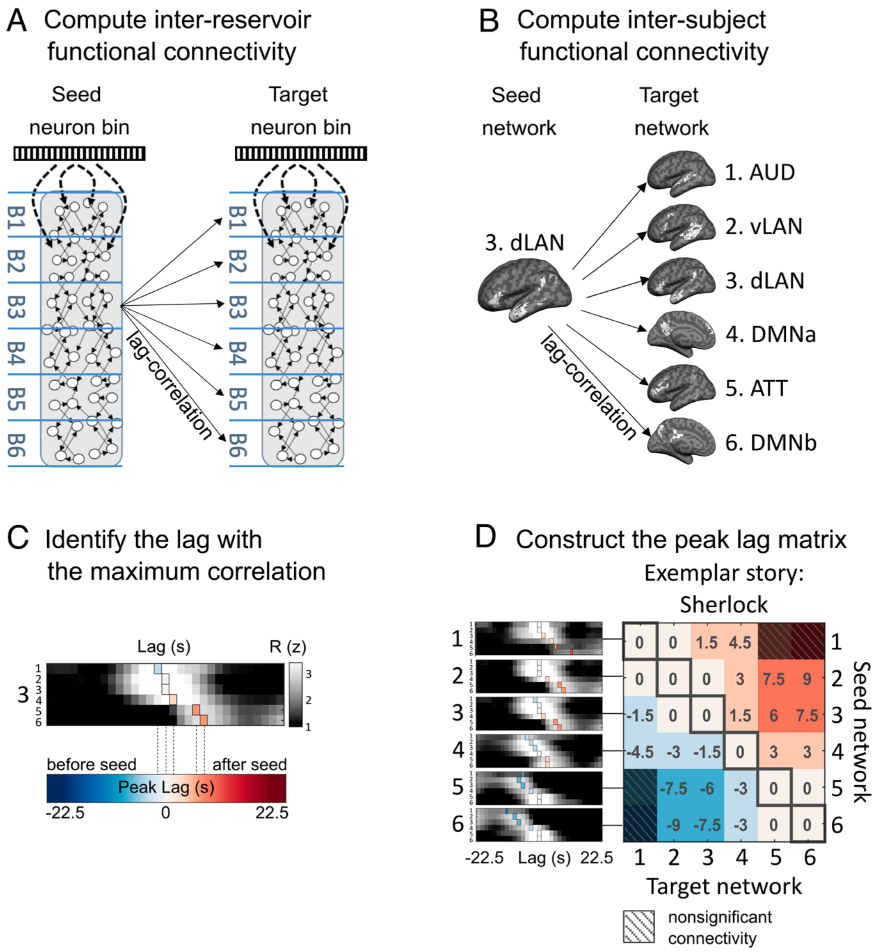

Fig. 1. Construction of the interregion peak lag matrix. (*A*) Lag-functional connectivity between seed-target neuron bins in reservoir networks, computed using an analysis method previously applied to the human brain (panels *B*–*D*, adapted from ref. 54). A network with limited input connectivity and canal-like connectivity (the Limited-Canal model) is shown as an example. Neurons are organized into bins, analogous to networks in panel *B*. (*B*) Lag-functional intersubject connectivities (cross-correlation) between seed-target brain networks were computed using the leave-one-subject-out method. Computing connectivity across subjects isolates stimulus-driven connectivity. The dorsal language (dLAN) network is used as an example seed network for illustrative purposes. AUD = auditory; vLAN = ventral language; ATT = attention; DMN = default mode network. (*C*) The matrix depicts intersubject functional connectivity between the dLAN seed and all six target networks at varying lags. The lag with the peak correlation value (colored vertical bars) was extracted and color-coded according to lag. For visualization, the lag-functional connectivities were z-scored across lags. (*D*) The network × network peak lag matrix ($P < 0.05$, false discovery rate (FDR) corrected) (55). Zeros along the diagonal indicate that intranetwork connectivity peaks at lag 0. Warm colors represent peak lags following the seed network, while cool colors represent peak lags preceding the seed network. An example story ("Sherlock") is shown for illustrative purposes.

in embedding space). The semantic structure of the narratives is not fully continuous over time: Narrative boundaries can emerge from semantic discontinuities in word-by-word embeddings derived from the narrative texts. For example, certain scenes or events will be semantically coherent, with high similarity among the corresponding word embeddings; at narrative event boundaries, there will tend to be a discontinuity in the similarity of the corresponding word embeddings (*SI Appendix*, Fig. S3). These discontinuities can occur at varying scales and provide the narrative boundaries that propagate through the network, driving the lag gradient. The word embeddings derived from the narrative text are supplied as input to the reservoir network.

Given this structured input, we systematically manipulated two structural parameters of the reservoir network: First, we manipulated whether external inputs are restricted to a subset of neurons, thus effectively focusing the input-driven activation on a small set of "sensory" neurons. Second, we manipulated the relationship between the probability that two neurons are connected and their topological nearness, such that neurons tend to communicate only with their closest neighbors. This connectivity architecture, often characterized by an exponential distance rule (EDR)—the probability of two neurons being connected decreases exponentially with the distance between them—is a characteristic of primate cortical connections (56). It has been shown to produce temporal processing hierarchies in models of the macaque cortex (3, 57).

To explore the influence of stimulus input distribution and connectivity architecture, we implemented four reservoir network models (Fig. 2). The Distributed-Random structure represents a classical reservoir model where stimulus inputs are distributed across all reservoir neurons, and reservoir neurons have an equal random probability of being connected to each other. In contrast, the Limited-Random model restricts input to a subset of sensory neurons while maintaining random connectivity. The Distributed-Canal model introduces a local connectivity rule while retaining distributed input. Finally, the Limited-Canal model combines both localized connectivity and constrained input, forming a structured activation flow that propagates from sensory neurons to more distant regions. These models allow us to investigate how different architectures influence information flow and temporal dynamics in reservoir networks. Among the candidate models, the reservoir network with limited input and canal-like connectivity architecture (Figs. 1*A* and 2*D*) most closely reflects the distribution of sensory neurons and anatomical connectivity in the human brain. We find that, in this model, narrative

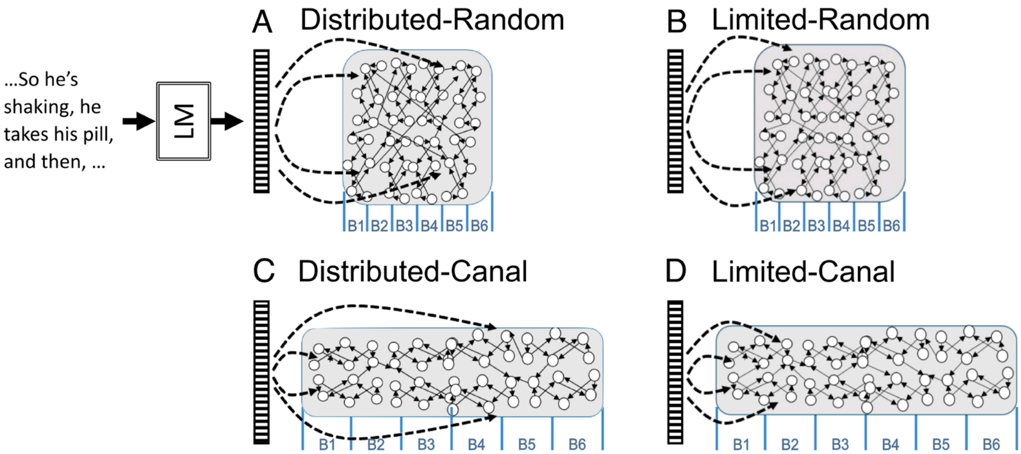

Fig. 2. Reservoir network with different structures. Narrative text is input to a language model (LM) which produces word embedding vectors that are input to the reservoir to simulate narrative processing. (*A*) Distributed-Random model (classic reservoir structure): Input is spread across all reservoir neurons, with random, uniform connectivity. (*B*) Limited-Random model: input is restricted to a subset of sensory neurons, while connectivity remains random. (*C*) Distributed-Canal model: Input is distributed, but connectivity follows a local rule. (*D*) Limited-Canal model: Both input and connectivity are constrained, resulting in a canal-like activation flow from sensory neurons to more distant areas. B1-B6 indicates six bins of consecutive neurons that make up the virtual brain regions for investigating the lag gradient. Adapted from ref. 52.

inputs reveal increasing temporal integration time constants along topographically organized neurons as found in humans (52).

To investigate the interaction between network structure and input structure, we included both intact natural narratives and scrambled versions of these narratives. The word-scrambled versions were created by shuffling the words, thereby eliminating relationships between words while keeping the individual words intact. The embedding-scrambled versions were created by shuffling the embeddings within each word, which, in addition to disrupting the relationships between words, further eliminates lexical semantics. The least structured embedding-scrambled versions serve as probes of the inherent properties of the reservoir models.

As predicted, the Limited-Canal setting endows the reservoir network with intrinsic lag properties (Fig. 3). More importantly, we found that the narrative structure enhances these inherent lag properties (Fig. 4). This sensitivity to input structure can be adjusted through the canal "width" parameter (Fig. 5), which regulates the maximum connection lengths. Namely, the same Limited-Canal architecture can be optimized to encode the hierarchical structure of naturalistic inputs. Finally, we show that processing cost, as a computational proxy for BOLD signals, builds up slower after event onset in higher-level regions (Fig. 6), which can account for the emergence of the lag gradient.

Results

We examined the dynamics of narrative-driven functional connectivity in reservoir networks varying along two structural parameters: 1) sensory inputs are either distributed across all neurons (see Fig. 2 *A* and *C*, dashed input line) or limited to a subset of neurons (300 out of 1,000 neurons; see Fig. 2 *B* and *D* dashed input line); and 2) neurons are randomly uniformly connected to other neurons (Fig. 2 *A* and *B*) or neurons are connected to neighboring neurons in a canal-like EDR fashion (Fig. 2 *C* and *D*). The consecutive neurons were binned into six virtual regions of 166 to 167 neurons, as illustrated in Fig. 2, for the models with limited sensory neurons (the Limited models, with inputs to neurons 1 to 300), only the two early regions receive sensory input directly. Furthermore, for the models created by the EDR (the Canal models), similar to the human brain, high-level regions receive

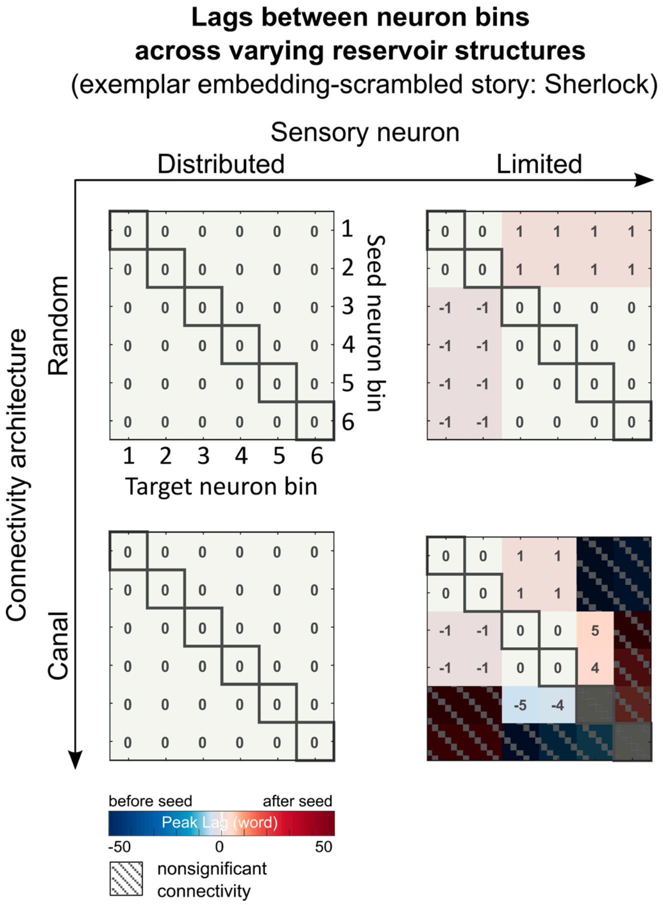

Fig. 3. The intrinsic lag gradient between neuron bins in reservoir networks of varying structure. We constructed reservoir networks with distributed or limited sensory neurons and random or canal-like connectivity structures. These models were then supplied with a sequence of scrambled embeddings from a naturalistic narrative (the Sherlock story in this example). Scrambled inputs were used to isolate the lag gradient induced by the intrinsic network architecture (as opposed to structure in the stimulus). We computed the input-driven functional connectivity between bins of neurons in the network across varying lags to quantify the lag gradient ($P < 0.01$, FDR correction). The Limited-Canal model replicates the lag gradient found in the human brain (54), reflecting the flow of information from early to later groups of neurons.

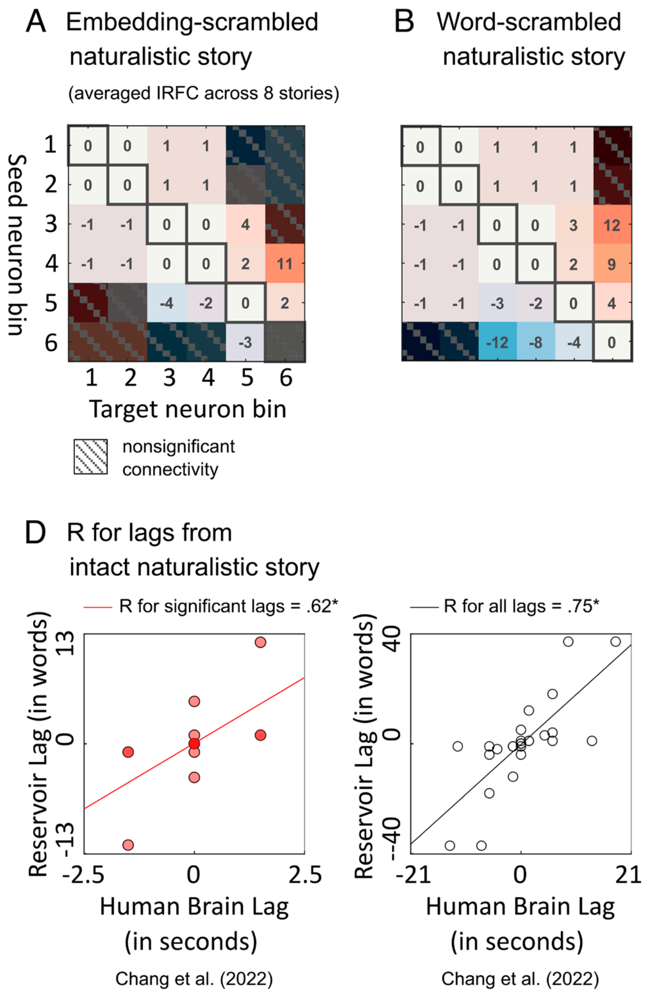

Fig. 4. Lag gradient in the Limited-Canal model is enhanced by narrative structure. We supplied embedding-scrambled, word-scrambled, and intact versions of eight naturalistic narratives from ref. 54 to the Limited-Canal reservoir and computed lagged functional connectivity. (*A*) Embedding-scrambled stories reveal the intrinsic lag properties of the Limited-Canal model. (*B*) A slightly stronger lag gradient was observed with the word-scrambled inputs. (*C*) Intact narratives induced the strongest lag gradient, in keeping with human results (54) ($P < 0.01$, FDR correction). (*D*) The scatter plot depicts the correlation between lag gradients in the human brain and the Limited-Canal model (*$P < 0.05$). Red dots indicate entries with significant lags in both reservoir and human peak lag matrices, while dark circles represent the remaining entries.

input-driven signals indirectly across a series of local connections (Fig. 2).

To investigate the interaction between network structure and input structure, we include intact, word-scrambled, and embedding-scrambled versions of the same narrative (*SI Appendix*, Fig. S3). Intact natural narratives are the most structured, and the embedding-scrambled version the least. To compare our findings with human results, we adopted the same naturalistic narratives and methods for estimating interregion lags as used in the original functional magnetic resonance imaging (fMRI) study (54).

While both strong positive and negative outputs from reservoir neurons indicate active engagement in the computation, when calculating the averaged signals within a neuron bin, positive and negative outputs from different neurons may cancel each other out, resulting in a flattened signal (*SI Appendix*, Fig. S4, *Upper*). To better approximate BOLD signal, which reflects hemodynamic changes driven by the metabolic demand of active neurons, following Hinaut and Dominey (58), we compute the absolute value of the change in activity between consecutive time steps for each neuron in the reservoir (*SI Appendix*, Fig. S4, *Lower*). This metric captures input-driven updates across neurons within a bin.

We measure functional connectivity between the six virtual regions by computing the Pearson correlation between their processing costs at varying lags (Fig. 1*A*). Interregion lags are estimated by extracting the lag, showing the maximal correlation value within the window of lags from −50 to +50 time steps (=50 words).

Intrinsic Lag Properties across Varying Reservoir Structures. We manipulated the distribution of sensory neurons (Distributed vs. Limited) and the reservoir network's architecture (Random vs. Canal). To examine the inherent lag properties of these models, we feed them with embedding-scrambled narratives, which are less structured than word-scrambled and intact narratives.

As predicted, we observed a gradient of lags in functional connectivity along bins of neurons—analogous to cortical areas—in the Limited-Canal model (Fig. 3). In this model, the external signals enter the network through a subset of sensory neurons and propagate progressively from early neurons to later neurons, like a canal with pebbles that can only drop at one end of it, causing waves to propagate over space and time toward the other end (Fig. 3*B*). This Limited-Canal model has topographically organized temporal integration constants and event representations of increasing durations (52) reproduced in *SI Appendix*, Figs. S1 and S2, supporting the claim by Chang et al. (54) that the narrative-driven lag gradient reflects the flow of information along the cortical hierarchy of processing timescales.

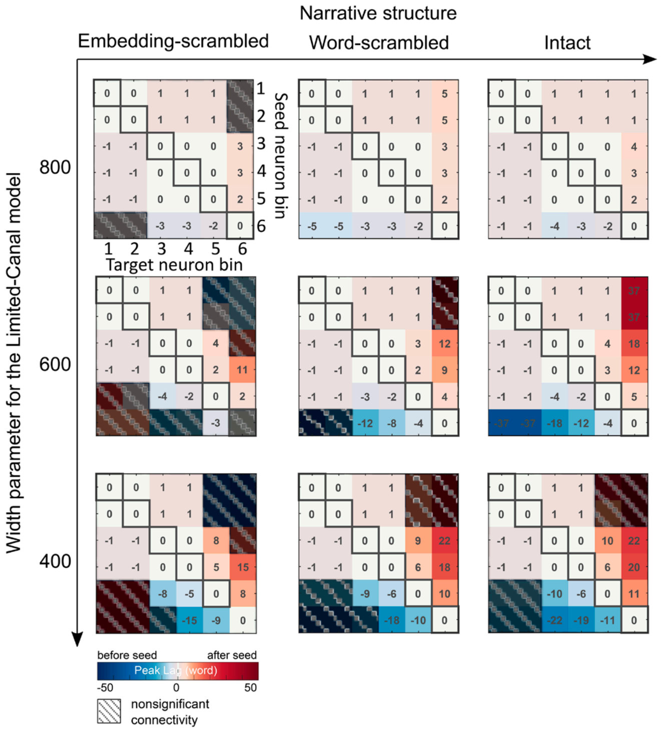

Fig. 5. The lag gradient in Limited-Canal models with varying canal widths. We varied the width parameter of the Limited-Canal model from 800 to 600 to 400, where a width of 800 trends toward fully random connectivity and 400 represents a more extreme "hose"-like canal. We evaluated the lag gradients in each model variation for embedding-scrambled, word-scrambled, and intact versions of eight naturalistic narratives from ref. 54 ($P < 0.01$, FDR correction). The Limited-Canal model with a width parameter set to 600 replicated the enhanced lag gradient with intact narratives found in the human brain (54). With a wider (800) architecture, the lag gradient was diminished, and the intact and scrambled yielded a similar (lack of) gradient. With a smaller width parameter (400), both the intact and scrambled narratives elicited a strong lag gradient, suggesting that the smaller width parameter imposes a lag gradient regardless of the structure of the input. Only the intermediate-width model yielded a strong lag gradient for the more structured intact stimulus and a weak gradient for the scrambled stimulus, in correspondence with the human brain.

On the other hand, the Limited-Random model shows a lag of only one time-step (i.e., one word) between bins 1 and 2 and the other bins. In this model, the one-word lag reflects the temporal lag between direct external input and the indirect impact of external input in neurons connected to the sensory neurons. The Distributed-Random and Distributed-Canal models do not show any lags between neuron bins. With distributed sensory neurons, the shared external input synchronizes different neuron bins and eliminates any activation lags between them. This also explains the synchronization between bin 1 and bin 2 in models with limited sensory neurons since only these two bins receive external inputs. Similarly, when dividing the 1,000 neurons into 20 (rather than six) bins, the first six bins, consisting of the first 300 neurons, are synchronized (*SI Appendix*, Fig. S5), due to the shared external inputs.

When fed with an intact narrative, the Limited-Canal model exhibits a stronger lag gradient, whereas the other models yield identical results with the embedding-scrambled narratives (*SI Appendix*, Fig. S6).

Narrative Structure Enhances the Intrinsic Lag Gradient. To better compare our results with findings in the human brain, we expanded our analysis to the same eight naturalistic narratives used by Chang et al. (54) to the Limited-Canal model. As shown in Fig. 4 *C* and *D*, we successfully replicated the lag gradient for the intact stories. Chang et al. (54) reported lags ranging from 0 to 9 s in human brain activity measured using fMRI. The Limited-Canal reservoir exhibits interarea lags ranging from 0 to 37 time-steps, approximately corresponding to 0 to 10 s (the reservoir receives one word per time step, and the mean word duration in Sherlock is 0.277 s; 37 time-steps × 0.277 s = 10.249 s).

More importantly, similar to the human results, scrambling the words in the narratives diminished the lag gradient in our model (Fig. 4*B*). Further scrambling at the embedding level not only disrupts word relationships but also removes lexical information, leading to an even weaker lag gradient (Fig. 4*A*). Providing fully random embedding values abolishes the lag gradient even further by removing information from the initial nonuniform distribution of embedding values (*SI Appendix*, Fig. S7).

Impact of Canal Width on the Reservoir Network's Sensitivity to Input Structure. We adjusted the canal width parameter to examine how lag properties intrinsic to the network structure interact with the narrative structure of the external stimulus. The larger this parameter, the longer the maximum length of connections, and thus, the more closely the network resembles the random network. This parameter was previously set to 600 for the preceding analyses and also for subsequent analyses unless otherwise specified.

Like the human brain, the Limited-Canal model with a width parameter set to 600 demonstrates sensitivity to input structure, exhibiting a stronger lag gradient with more structured inputs (Figs. 3 and 4). The strongest lag gradient was observed with intact narratives, while embedding-scrambled narratives induced the weakest lag gradient, reflecting the intrinsic lag properties of this model. On the other hand, neural dynamics in reservoirs with stronger or weaker inherent lag properties fail to reflect narrative structure (Fig. 5). With a larger width parameter (800), we observed a diminished lag gradient even with an intact story because the longer connections allow more synchronizing signals to be shared between neuron bins. With a narrower width parameter (400), a strong lag gradient is observed even with scrambled narratives. As the canal is narrower and longer, intuitively, the canal has become a hose; intrinsic lag properties due to extreme local structural connectivity dominate input-driven functional connectivity. These simulation results suggest that the reservoir network's sensitivity to input structure can be modulated by its structural parameter.

Processing Costs Build up More Slowly in Later Neurons. Next, we investigated the mechanism underlying the lag gradient. Our previous simulation suggests that the lag gradient emerges from information integration at different nested granularities (e.g., word, phrase, sentence, etc.), which leads to a slower accumulation of activity in higher-level brain areas (54). We examined the processing cost around event boundaries when information integration resets (4) in the reservoir. We created eight synthetic narratives with known event boundaries. By feeding synthetic narratives to the reservoir, we can examine the pure semantic boundary effect with exact knowledge of where the event boundary occurs. We computed each neuron bin's average processing cost across events and stories.

For each synthetic narrative, we computed the average processing cost around event boundaries. To measure the 50% peak latency, we first identified the maximum processing cost within 0 to 50 words after event onset. The peak latency was then defined as the first time point at which the processing cost reached 50% of this peak amplitude (*SI Appendix*, Fig. S8). See *SI Appendix*, Fig. S9 for processing costs in naturalistic narratives.

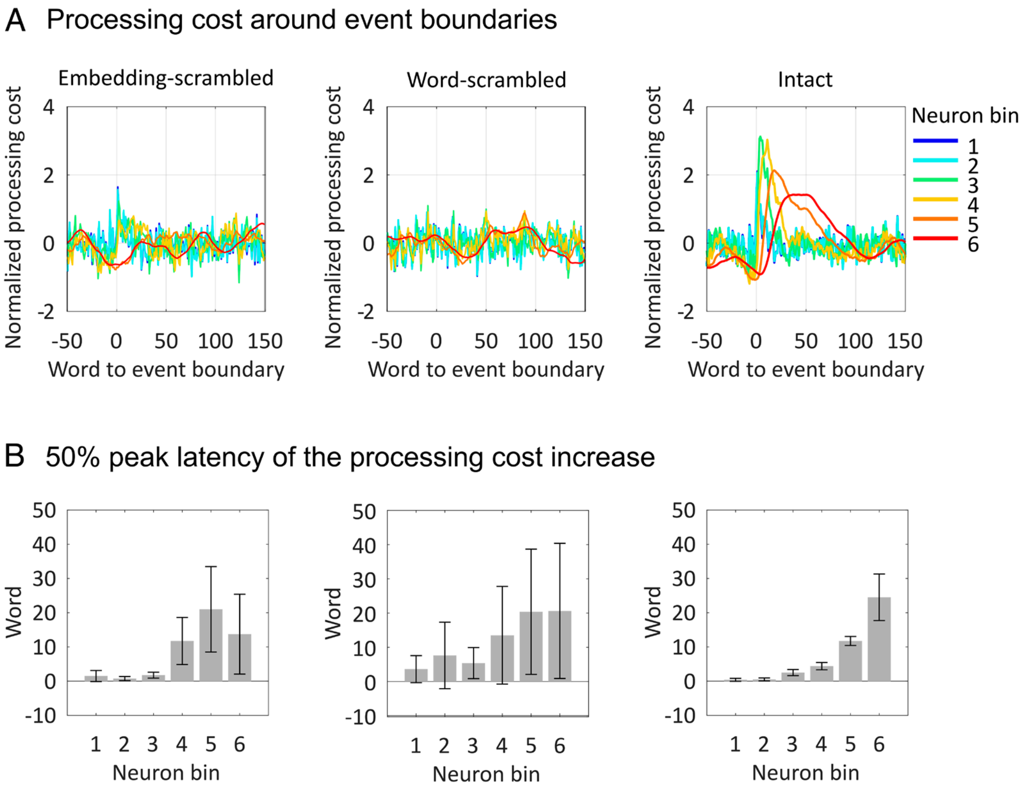

Fig. 6. Processing cost around event boundaries. (*A*) We fed intact, word-scrambled, and embedding-scrambled synthetic narratives into the Limited-Canal model (width = 600) and extracted processing costs around event boundaries. Since the word-scrambled version lacks inherent event boundaries, we used those from the intact version. The averaged, z-scored time courses are shown. (*B*) 50% peak latencies of processing cost increases after event onsets (*SI Appendix*, Fig. S8). Error bars show 95% CI across stories. Under the intact conditions, latency differences were significant between all neuron bin pairs except between bin 1 and bin 2 (both containing sensory neurons) (two-sample *t* test, df = 7, $P < 0.05$, corrected for familywise error (FWE)). Processing costs rising more slowly and persisting longer in later neurons, especially in intact narratives, may contribute to the lag gradient. No significant latency differences were found between neuron bins in the two scrambled conditions after FWE correction.

Similar to naturalistic narratives, we found an enhanced lag gradient in the intact synthetic narratives (*SI Appendix*, Fig. S7). As shown in Fig. 6, processing costs increase after event onsets, and more importantly, processing costs build up more slowly in later neurons, as suggested by Chang et al. (54). In coherent narratives, successive words tend to exhibit semantic associations and correlations among embeddings. Our results indicate that semantic shifts at event boundaries leads to an increase in processing costs, which propagates throughout the network. Scrambling words disrupts the event structure, thus weakening the lag gradient.

Discussion

Our results show that the reservoir network with most brain-like anatomical topology (i.e., Limited-Canal model) (Fig. 2*D*), best captured the sensitivity to narrative structure observed in the neural dynamics of the human brain (Figs. 3 and 4). This finding suggests that while the same anatomical structure can produce different functional dynamics (59, 60), depending on the current task and prior training, the lag gradient at least partially arises from the network connectivity characterized by hierarchical structures. Specifically, we endowed the structure of a reservoir network with local connectivity and restricted input to a limited set of sensory neurons to replicate a gradient of lagged connectivity between stages of cortical processing hierarchy, as observed in humans listening to natural spoken narratives (54). When we fed minimally structured inputs, namely, embedding-scrambled narratives into these reservoir models, we observed that reservoirs with limited sensory neurons and a canal-like local connectivity structure already exhibited intrinsic lag properties (Fig. 3). Importantly, we demonstrated that this intrinsic lag gradient strengthens with inputs containing more internal structure, and is strongest with intact narratives (Fig. 4). This sensitivity to input structure could be adjusted by manipulating the maximum length of connectivity between neurons in the Limited-Canal model, namely, the canal "width" parameter. We found that human-like neural dynamics, namely, stronger lag gradients with more structured inputs, emerged only under intermediate canal widths (Fig. 5). There seems to be an optimal architecture which allows the network to encode information in a way that highlights the intrinsic hierarchical structure of the naturalistic input. Finally, using synthetic narratives with known semantic event boundaries, we showed that processing costs at event boundaries accrue more slowly in later neurons (Fig. 6). This finding suggests a potential mechanism underlying the emergence of the observed lag gradient.

In the Limited-Canal model, activation from input-driven neurons must propagate through local connections to reach neurons located progressively farther away. Intuitively, these neurons can be viewed as a succession of low pass filters where successive (higher order) neurons thus respond with progressively slower dynamics. As shown in Fig. 6, neurons that are farther from the input source exhibit increasing response delays, and their activity persists for longer durations. Critically, because all neurons share the same leak rate, which determines how much of each single neuron's previous state is retained at each time step (see the *Materials and Methods* section for details), these additional delays must arise from macroscale architectural features of the network, namely the input distribution and local connectivity.

To gain a more mechanistic understanding of the network's temporal dynamics, we introduced brief pulse inputs to each neuron bin in turn and examined their temporal response dynamics. For each neuron, we measured the maximum value of its response to the input pulse, and then identified the time when its response returned to 20% of this value, as the impulse response time constant. As shown in *SI Appendix*, Fig. S10, the bin directly receiving the input consistently exhibited the shortest time constant, indicating tight coupling to fast-changing inputs. In contrast, time constants increased progressively in bins located farther from the input-driven bin. These results demonstrate that the temporal dynamics of each bin are not determined by its intrinsic properties, but are shaped by its topographical position relative to the input source.

An intrinsic lag gradient has been reported during the resting state (60–62) and with scrambled narrative stimuli (54, *SI Appendix*, Fig. S8*B*) in the human brain using within-subject functional connectivity, although at a scale much smaller (−1 to 1 s) than the narrative-driven lag gradient (up to 9 s). Similar to the human brain, the Limited-Canal model shows intrinsic lag properties with embedding-scrambled narratives (Fig. 3), which are enhanced by narrative structure (Fig. 4) with a width parameter set to 600 (Fig. 4). The small interregion lag in the wider model (width parameter 800) for both intact and scrambled narratives supports the claim that intrinsic lag properties are necessary to simulate human results. A strong lag gradient in the narrower model (width parameter 400), even under the scrambled condition, shows that not all network structures with intrinsic lag properties can reproduce the human results; in this network, the network structure imposes the lag gradient regardless of the input structure. The intermediate model (width parameter 600), in which input signals propagate to but do not dominate high-level neurons, behaves most similarly to the human brain.

Note that, in the human studies, the intrinsic lag gradient was only observed with within-subject functional connectivity analysis (54, 60–62), which retains idiosyncratic intrinsic signals, while the narrative-driven lag gradient was revealed only by intersubject functional connectivity analysis (54), which isolates the input-driven signals (63, 64). Unlike the human brain, the reservoir network does not have intrinsic activity fluctuations independent of external input, which explains why interreservoir and within-reservoir analyses yield similar results (*SI Appendix*, Fig. S11).

In our previous simulation (54), different levels of structural boundaries between events and their temporal integration functions were explicitly specified. In contrast, this study provides word embeddings from naturalistic narratives, which contain inherent semantic structures but no explicitly assigned boundaries. Moreover, the reservoir models' responses to these implicit semantic boundaries are not governed by a predefined temporal integration function but the models' network structures. We found that semantic discontinuities at event boundaries result in abrupt shifts in the geometric orientations of word embeddings when transitioning to a new event (51). This leads to increased processing costs that propagate through the network (Fig. 6). Word scrambling can effectively remove these processing cost increases because it disrupts local semantic coherence (corresponding to events), thereby preventing the propagation of processing costs through the network and reducing the lag gradient (Fig. 4*B*).

Chang et al. (54) showed that the lag gradient is robust to varying lengths of linguistic/narrative units (for example, sentence length) and speech rate, as well as some variations in the temporal integration function, but is impacted by interparagraph silences, which naturally occur in spoken narratives. Processing cost after event onset (Fig. 6), as an analog to the temporal integration function, does not exactly correspond in shape to the linearly increasing temporal integration function adopted by Chang et al. (54). This may explain why interparagraph silent pauses are not necessary for the emergence of the lag gradient from our model. Unlike the BOLD signals simulated by Chang et al. (54), processing costs started to decrease before event offsets, probably because the novelty of the synthetic input saturates by the end of events. Therefore, pauses without any input are not necessary to desaturate the processing cost.

A hierarchy of processing timescales is suggested to underlie both the lag gradient and the covarying lengths of cortical events revealed by event segmentation analyses (54). Indeed, in the Limited-Canal model, we observed both phenomena (Fig. 4 and *SI Appendix*, Fig. S2). Areas farther from the input display a greater lag compared to closer areas, and they also exhibit a preference for longer events as quantified using hidden Markov models (HMMs). This suggests that the lag gradient and increasingly longer-duration events are intrinsically related. Interestingly, this is not the case for sorted neurons in the Distributed-Random [the classic model in Dominey (51)]. While we were able to confirm that slower neurons prefer longer events, we did not observe the lag gradient in the sorted neurons in the classic model (*SI Appendix*, Fig. S12). This is because the classic model violates one of the structural conditions required to produce the lag gradient: Input activations must be restricted to a subset of neurons. Because the inputs are projected with a fixed probability to all reservoir neurons in the classic model, these neurons are synchronized by the common input and thus cannot produce the lag gradient. These results show that the lag gradient provides a complementary insight relative to the structure revealed by the HMM.

We found that the model with the more brain-like structure (i.e., Limited-Canal model) (Fig. 2*D*) best reproduced neural dynamics observed in the human brain (Figs. 3 and 4). In a similar vein, it has been reported that more biologically realistic architectures maximize the reservoir network's memory capacity relative to the wiring cost (48). One way to further improve the Limited-Canal model is to introduce a long-distance shortcut as an analog to white matter fasciculi from input-driven sensory areas to frontal associative areas that are quite far apart in terms of Euclidean distance. For example, the inferior fronto-occipital fasciculus (65) provides a direct shortcut between the occipital and frontal cortex. This could produce inconsistencies in the temporal processing hierarchy. Interestingly, such inconsistencies can be observed in Chien and Honey (4). In their data, a medial prefrontal area in the DMN (LH_DefaultB_PFCd_1) has an unusually fast integration time constant relative to its anatomical location. We suggested that this is due to long-distance white matter connections that provide a direct path from input-driven areas to this region. We demonstrated how such effects could be produced in simulations with the reservoir model (50, 52). Future research should determine the effects of these long-distance connections on the lag gradient.

In conclusion, we found that the brain-inspired Limited-Canal architecture (Fig. 2*D*) endows the recurrent reservoir network with intrinsic lag properties even without task-specific training (Fig. 3). Interestingly, while scrambled input reveals an intrinsic lag gradient, structured semantic input from narrative word embeddings further enhances the lag gradient (Fig. 4). This sensitivity to input structure is modulated by the structural parameter, canal width (Fig. 5). These findings imply a "sweet spot" when adapting the network structure to the structure of naturalistic input. Similarly, the structure of the human brain may be optimized through evolution and learning to process hierarchical regularities in the environment (66). The brain, in turn, generates language/narratives with an isomorphic structure (53, 67), where different levels of temporal structure in narrative may have evolved to match the representational capabilities of the cortex (51).

Materials and Methods

Reservoir Networks. In reservoir computing, a random dynamic recurrent neural network is stimulated with input, and the resulting rich high dimensional states are then harvested (44). Typically this harvesting consists in training the output weights from reservoir units to output units, and then running the system on new inputs and collecting the resulting outputs from the trained system. In the current research, we focus our analysis directly on the rich high-dimensional states in the reservoir itself. That is, we do not train the reservoir to perform any transformation on the inputs. Instead, we analyze the activity of the reservoir neurons themselves, as a proxy for cortical activity. Neural activity in such recurrent networks has been demonstrated to be usefully comparable to primate cortical activity (47, 68, 69).

The reservoir simulation of human narrative processing consists of two components. First, the language model (LM) generates word embedding vectors, and second, the reservoir generates from these embeddings the spatiotemporal trajectory of neural activation. Given the input narrative, input words are transformed into 100-dimensional word embedding vectors by the Wikipedia2Vec model, pretrained on the three billion words 2018 Wikipedia corpus (70). These vectors are then input to the reservoir, a neural network with fixed recurrent connections. Because of the human narratives we use, and the LM used to generate embeddings, this input is already a structured trajectory, and the reservoir's internal states encode and generate a rich high dimensional representation of this trajectory (51).

Our discrete-time, tanh-unit echo state network with N reservoir units and K input units is characterized by the state update equation:

$$x(t+1) \;=\; (1-\alpha)\mathrm{x}(\mathrm{t}) + \alpha \cdot f\left(\mathbf{W}\mathrm{x}(\mathrm{t}) + \mathrm{W}_{\mathrm{in}}\mathrm{u}(\mathrm{t})\right), \qquad [1]$$

where x(n) is the N-dimensional reservoir state, f is the tanh function, W is the N × N reservoir weight matrix, W_{in} is the N × K input weight matrix, u(n) is the K dimensional input signal, α is the leak rate (0.2). The matrix elements of W and W_{in} are drawn from a random distribution.

The reservoir was instantiated using easyesn, a python library for recurrent neural networks using echo state networks (https://pypi.org/project/easyesn/) (71). We used a reservoir of N = 1,000 neurons, with input dimension of K = 100. The W and W_{in} matrices are initialized with uniform distribution of values from −0.5 to 0.5. Multiple reservoir instances (corresponding to experimental subjects) were generated by using different seed values in this initialization. Unless otherwise stated, we used 40 distinct reservoir instances in each of the different experiments. The leak rate was 0.2, a standard value that allows the neuron to maintain an influence of its history and sensitivity to inputs. In order to implement the connectivity architecture where EDR applies, i.e. the canal models, we applied the following procedure to the connection weights W. All connections greater than max_length (600), are set to zero. Then, for all connections W[i,j] of length equal abs(i−j):

$$\begin{aligned}\mathrm{W}[\mathrm{i},\mathrm{j}] \;=\; &\mathrm{W}[\mathrm{i},\mathrm{j}] * \left(\left(\mathrm{max\_length} - \mathrm{length}\right)/\mathrm{max\_length}\right)^{3} \\ &* \left(1 + \mathrm{i} * \mathrm{gradient}\right) * \mathrm{gain}.\end{aligned} \qquad [2]$$

The parameter gradient = 7.5e−4 and gain = 1.75. The gradient term provides for an increase in local connectivity as the distance from the input increases, as observed in the cortex (57). The max_length parameter was 600, except when we specifically tested the effects of varying it to 400 or 800. For the models with limited sensory neurons, the input matrix was accordingly updated so that inputs from the 100-element word embedding vectors were provided to only the first 300 (bins 1 and 2) of the 1,000 neurons (Fig. 7).

In summary, in order to establish the structural requirements for reservoirs that are necessary to produce the lag gradient, we examine reservoirs that vary according to two parameters, the distribution of sensory neurons and the reservoir architecture. Sensory neurons, i.e., neurons receiving external input, are either distributed over the entire reservoir (distributed) or limited to a subset of neurons (limited). The architecture refers to whether the reservoir itself has a connectivity structure with neurons being connected with uniform probability (random), or having the probability of connectivity between neurons proportional to the distance between them, thus producing a flow along these local connections (canal). Canal width refers to the maximum length of connections, as expressed in Eq. **2**. The majority of the experiments are performed with canal width of 600, with specific tests to examine the effects of width.

Narrative Inputs.

Naturalistic narratives. Eight naturalistic narratives were directly generated from the text transcripts of the narratives used by Chang et al. (54). The transcripts correspond to the narratives: Sherlock, "Merlin," "The 21st year," "Pie Man (PNI)," "I Knew You Were Black," "The Man Who Forgot Ray Bradbury," "Running from the Bronx (PNI)," and "Pie Man" (72). The transcripts are available at http://datasets.datalad.org/?dir=/labs/hasson/narratives/stimuli/transcripts. Each word in the text file was used to generate a 100-element word embedding vector using Wikipedia2Vec. Rare (<4%) words that were not found in the three billion word corpus were skipped. The eight resulting narrative inputs varied in length from 947 to 2861 words (mean = 1,798 words).

Synthetic narratives. Eight synthetic narratives were generated from randomly selected Wikipedia pages, using https://pypi.org/project/wikipedia/. Each narrative was made up of 20 events. A given Wikipedia page was used to generate one narrative event, meaning that each event revolved around the topic of that Wikipedia page and was generally unrelated to the other 19 events. A given event had a minimum of 25 words, plus a variable component that varied from 0 to 150 words, for a mean event length of 100 words. Once the narrative was thus generated, each word was used to generate the 100-element word embedding. The resulting eight narratives varied in length from 1,753 to 2,227 words (mean = 1,948 words). We generated these synthetic narratives in order to have direct access to the ground truth for event boundaries.

Word scrambling. Each narrative is represented as a sequence of the 100-element word embeddings. In the word scrambling process, the order of the sequence was permuted, while the embeddings were maintained intact. This manipulation preserves the overall structure of the embeddings themselves, but will tend to disrupt the narrative structure of the embedding sequence.

Embedding scrambling. In embedding scrambling, each of the 100-element word embeddings in the sequence of embeddings was permuted, each with a different permutation. This disrupts any regularities in embedding dimensions across words (as well as the overall narrative structure), while preserving the same general distribution of embedding values.

Reservoir Data Preprocessing.

Processing cost. In the reservoir network, the raw output of each neuron ranges from –1 to 1, resulting from the random initialization of input and reservoir weights between –0.5 and 0.5, with tanh as the activation function. Both positive and negative activations indicate neural engagement in computation. Since BOLD signal reflects hemodynamic changes driven by the metabolic demand of active neurons, following Hinaut and Dominey (58), we computed the absolute change in activity between consecutive time steps for each neuron. This metric captures input-driven updates across neurons within a bin and aligns with the idea that integrating new words becomes less effortful when they are more predictable, such as toward the end of a sentence, due to the accumulation of contextual information (73–75).

Virtual region of interest. We create six virtual regions of interest (ROI) by binning the neurons into groups, each consisting of 166 to 167 neurons. For the models with limited sensory neurons (Limited-Random and Limited-Canal models), only the two early ROIs receive input directly. For the models created by the EDR (Limited-Canal and Distributed-Canal models), the virtual ROIs reflect the topology of the network. Namely, high-level regions could receive input-driven signals indirectly across a series of local connections.

Interreservoir Functional Connectivity (IRFC). Chang et al. (54) computed intersubject functional connectivity at different lags to estimate the temporal lag between neural networks. In this study, we compute IRFC between virtual areas, using the leave-one-out method; i.e., correlation between the time series from each reservoir instance and the average time series of all the other instances (63).

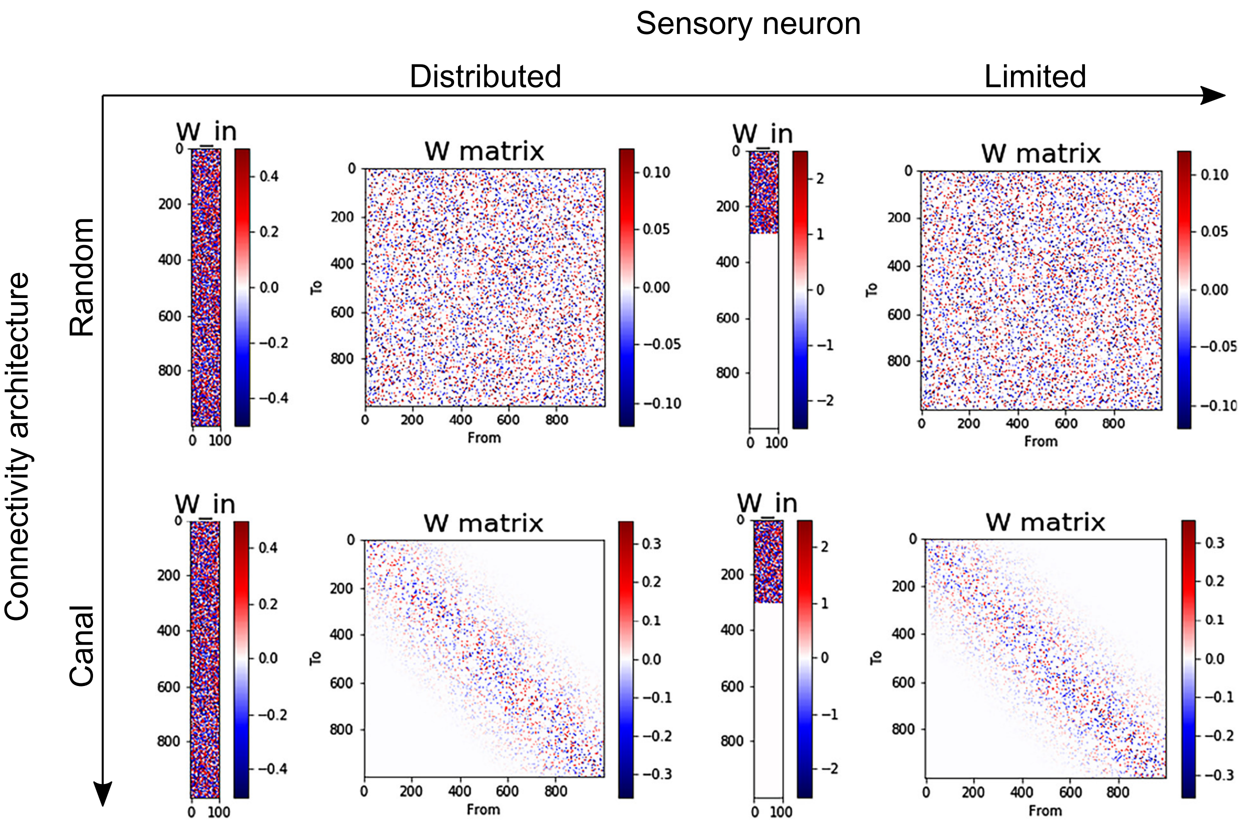

Fig. 7. Reservoir topologies. Reservoir topologies vary along two dimensions, the distribution of neurons receiving the input (distributed/limited) and whether reservoir connectivity is constrained by the EDR (no/random/ or yes/canal). The EDR produces a canal-like topology. (*Upper Left*) In the classic Distributed-Random configuration, the 100-element word embedding vector is projected to all reservoir neurons (W_in). The connectivity matrix specifies that all neurons project to all other neurons with a density of 0.2. (*Lower Right*) In the Limited-Canal reservoir, inputs are restricted to the first 300 reservoir units. The connectivity matrix is structured by an EDR whereby the probability of two neurons being connected decreases exponentially with their distance. This yields a connectivity matrix that is organized along the diagonal where i ~ =j in the W matrix. For the classic reservoir, neurons with different temporal integration constants are scattered across the network, while for the Limited-Canal reservoir, topologically organized integration time constants are clearly observed across successive groups of neurons with distance from the input (*SI Appendix*, Fig. S1).

Before computing the correlation, data from the first 400 and last 20 time steps were discarded to remove large signal fluctuations at the beginning and end of the time course due to signal stabilization and stimulus onset/offset. We then averaged time series across neurons within each bin and z-scored the resulting time series.

Lag-correlations were computed by circularly shifting the time series such that the nonoverlapping edge of the shifted time series was concatenated to the beginning or end. The left-out reservoir was shifted while the average time series of the other reservoir instances remained stationary. Fisher's z transformation was applied to the resulting correlation values prior to further statistical analysis.

Peak Lag Matrix. Following Chang et al. (54), we computed the region × region × lag-IRFC matrix and extracted the lag with peak correlation value for each region pair. The peak correlation value was defined as the maximal IRFC value within the window of lags from −50 to +50 time steps (=50 words). The mean word duration in Sherlock is approximately 0.277 s. Thus 50 time steps roughly correspond to 15 s; we required that the peak IRFC be larger than the absolute value of any negative peak and excluded any peaks occurring at the edge of the window.

To exclude IRFC peaks that only reflected shared spectral properties, we generated surrogates with the same mean and autocorrelation as the original time series by time-shifting and time-reversing. We computed the correlation between the original seed and time-reversed target with all possible time shifts. Since we shift the time series circularly, the number of possible lags equals the number of time points. The resulting IRFC values served as a null distribution. A one-tailed z-test was applied to compare IRFCs within the window of lag −50 to +50 TRs against this null distribution. The FDR method was used to control for multiple comparisons (seed × target × lags; $q < 0.01$) (55). When assessing IRFC for each story, only this first test was applied.

For the mean IRFC across stories, a second statistical test was also applied, i.e., a parametric one-tailed one-sample *t* test to compare the mean IRFC against zero ($N = 8$ real or synthetic stories) and corrected for multiple comparisons by controlling the false discovery rate (FDR; 6 seed × 6 target × 31 lags; $q < 0.01$) (55). Only IRFC that passed both tests were considered significant.

The Latency of Processing Cost after Event Onset. The processing costs around event boundaries (−50 ~ 100 time steps) were extracted and averaged across events and reservoirs for each of the eight synthetic stories. The latency of processing cost after event onset was defined as the time step when the processing cost reached 50% of its peak amplitude. Paired *t* tests ($N = 8$) were applied to compare processing cost latencies in different neuron bins. The FWE method was used to control for multiple comparisons (15 bin pairs).

Data, Materials, and Software Availability. This research is realized in the open code spirit, and indeed benefitted from open code and data for development of the reservoir model (71), and the language model for word embeddings. The original Narrative Integration Model code in python (51), and all required data are available on GitHub https://github.com/pfdominey/Narrative-Integration-Reservoir/. The EDR model used in this study, including all necessary data, can be found at: https://github.com/pfdominey/Reservoir_lag_gradient/.

ACKNOWLEDGMENTS. This study is supported by the National Institute of Mental Health, USA (R01-MH112357 and DP1-HD091948), the National Science and Technology Council, Taiwan (NSTC113-2410-H-038-031-MY2), Taipei Medical University (TMU112-AE1-B21), and the Ministry of Education in Taiwan (Higher Education Sprout Project) (DP2-TMU-113-N-03 and DP2-TMU-114-N-03).

1. J. B. Burt *et al.*, Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat. Neurosci.* **21**, 1251–1259 (2018).
2. J.-P. Changeux, A. Goulas, C. C. Hilgetag, A connectomic hypothesis for the hominization of the brain. *Cereb. Cortex* **31**, 2425–2449 (2021).
3. R. Chaudhuri, K. Knoblauch, M. A. Gariel, H. Kennedy, X.-J. Wang, A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron* **88**, 419–431 (2015).
4. H. Y. S. Chien, C. J. Honey, Constructing and forgetting temporal context in the human cerebral cortex. *Neuron* **106**, 675–686.e11 (2020).
5. L. Cocchi *et al.*, A hierarchy of timescales explains distinct effects of local inhibition of primary visual cortex and frontal eye fields. *Elife* **5**, e15252 (2016).
6. M. Demirtaş *et al.*, Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron* **101**, 1181–1194.e13 (2019).
7. J. M. Huntenburg, P. L. Bazin, D. S. Margulies, Large-scale gradients in human cortical organization. *Trends Cogn. Sci.* **22**, 21–31 (2018).
8. C. A. Runyan, E. Piasini, S. Panzeri, C. D. Harvey, Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).
9. L. Pylkkänen, The neural basis of combinatory syntax and semantics. *Science* **366**, 62–66 (2019).
10. U. Hasson, J. Chen, C. J. Honey, Hierarchical process memory: Memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
11. C. J. Honey *et al.*, Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).
12. Y. Lerner, C. J. Honey, L. J. Silbert, U. Hasson, Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
13. C. J. Honey, C. R. Thompson, Y. Lerner, U. Hasson, Not lost in translation: Neural responses shared across languages. *J. Neurosci.* **32**, 15277–15283 (2012).
14. M. Nguyen, T. Vanderwal, U. Hasson, Shared understanding of narratives is correlated with shared neural responses. *Neuroimage* **184**, 161–170 (2019).
15. M. Regev, C. J. Honey, E. Simony, U. Hasson, Selective and invariant neural responses to spoken and written narratives. *J. Neurosci.* **33**, 15978–15988 (2013).
16. A. Zadbood, J. Chen, Y. C. Leong, K. A. Norman, U. Hasson, How we transmit memories to other brains: Constructing shared neural representations via communication. *Cereb. Cortex* **27**, 4988–5000 (2017).
17. C. Baldassano *et al.*, Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).
18. C. Baldassano, U. Hasson, K. A. Norman, Representation of real-world event schemas during narrative perception. *J. Neurosci.* **38**, 9689–9699 (2018).
19. L. Geerligs *et al.*, A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. *Elife* **11**, e77430 (2022).
20. A. Truzzi, R. Cusack, The development of intrinsic timescales: A comparison between the neonate and adult brain. *Neuroimage* **275**, 120155 (2023).
21. T. S. Yates *et al.*, Neural event segmentation of continuous experience in human infants. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2200257119 (2022).
22. T. Hannagan, A. Agrawal, L. Cohen, S. Dehaene, Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104779118 (2021).
23. M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, F. P. Lange, A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2201968119 (2022).
24. T. C. Kietzmann *et al.*, Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21854–21863 (2019).
25. L. Kozachkov, K. V. Kastanenka, D. Krotov, Building transformers from neurons and astrocytes. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2219150120 (2023).
26. M. Schrimpf *et al.*, The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105646118 (2021).
27. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
28. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, Eds. (Curran Associates Inc., Red Hook, NY, 2013), pp. 3111–3119.
29. K. Clark, U. Khandelwal, O. Levy, C. D. Manning, "What does BERT look at? An analysis of BERT's attention" in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, , T. Linzen, G. Chrupała, Y. Belinkov, D. Hupkes, Eds. (Association for Computational Linguistics, Stroudsburg, PA, 2019), pp. 276–286.
30. I. Tenney, D. Das, E. Pavlick, "BERT rediscovers the classical NLP pipeline" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, , A. Korhonen, D. Traum, L. Màrquez, Eds. (Association for Computational Linguistics, Stroudsburg, PA, 2019), pp. 4593–4601.
31. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, J. Burstein, C. Doran, T. Solorio, Eds. (Association for Computational Linguistics, Stroudsburg, PA, 2019), pp 4171–4186.
32. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training". (OpenAI, San Francisco, CA, 2018).
33. A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates Inc., Red Hook, NY, 2017), pp. 5998–6008.
34. C. J. Caucheteux, A. Gramfort, J.-R. King, "Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects" in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, S. W. Yih, Eds. (Association for Computational Linguistics, Stroudsburg, PA, 2021), pp. 3635–3644.
35. C. Caucheteux, A. Gramfort, J.-R. King, Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
36. A. Goldstein *et al.*, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
37. S. Kumar *et al.*, Shared functional specialization in transformer-based language models and the human brain. *Nat. Commun.* **15**, 5523 (2024).
38. M. Toneva, L. Wehbe, "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates Inc., Red Hook, NY, 2019), pp 14954–14964.
39. M. H. Christiansen, N. Chater, The now-or-never bottleneck: A fundamental constraint on language. *Behav. Brain Sci.* **39**, e62 (2015).
40. J. L. Elman, Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
41. M. E. Peters, M. Neumann, L. Zettlemoyer, W. T. Yih, "Dissecting contextual word embeddings: Architecture and representation" in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii Eds. (Association for Computational Linguistics, Stroudsburg, PA, 2018), pp. 1499–1509.
42. A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
43. D. V. Buonomano, W. Maass, State-dependent computations: Spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10**, 113–125 (2009).
44. M. Lukoševičius, H. Jaeger, Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* **3**, 127–149 (2009).
45. M. Rigotti *et al.*, The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
46. N. T. Markov *et al.*, Weight consistency specifies regularities of macaque cortical networks. *Cereb. Cortex N.Y.* **21**, 1254–1272 (2011).
47. P. Enel, E. Procyk, R. Quilodran, P. F. Dominey, Reservoir computing properties of neural dynamics in prefrontal cortex. *PLoS Comput. Biol.* **12**, 455–462 (2016).
48. L. E. Suárez, B. A. Richards, G. Lajoie, B. Misic, Learning function from structure in neuromorphic networks. *Nat. Mach. Intell.* **3**, 771–786 (2021).
49. L. E. Suárez *et al.*, Connectome-based reservoir computing with the conn2res toolbox. *Nat. Commun.* **15**, 656 (2024).
50. P. Triebkorn, V. Jirsa, P. F. Dominey, Simulating the impact of white matter connectivity on processing time scales using brain network models. *Commun. Biol.* **8**, 1–12 (2025).
51. P. F. Dominey, Narrative event segmentation in the cortical reservoir. *PLoS Comput. Biol.* **17**, e1008993 (2021).
52. P. F. Dominey, T. M. Ellmore, J. Ventre-Dominey, "Effects of connectivity on narrative temporal processing in structured reservoir computing" in *Proceedings of the International Joint Conference on Neural Networks*, M. Gori, A. Sperduti, Eds. (IEEE, Piscataway, NJ, 2022), pp. 1–8.
53. S. J. Kiebel, J. Daunizeau, K. J. Friston, A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* **4**, e1000209 (2008).
54. C. H. C. Chang, S. A. Nastase, U. Hasson, Information flow across the cortical timescale hierarchy during narrative construction. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209307119 (2022).
55. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
56. M. Ercsey-Ravasz *et al.*, A predictive network model of cerebral cortical connectivity based on a distance rule. *Neuron* **80**, 184–197 (2013).
57. R. Chaudhuri, A. Bernacchia, X.-J. Wang, A diversity of localized timescales in network activity. *Elife* **3**, e01239 (2014).
58. X. Hinaut, P. F. Dominey, Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PLoS One* **8**, e52946 (2013).
59. G. J. Cooper, G. Blackburne, T. M. Dekker, R. K. Das, J. I. Skipper, Where the present gets remembered: Sensory regions communicate with the brain over the longest timescales. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.09.18.558347v2 (Accessed 20 September 2023).
60. A. Mitra, A. Z. Snyder, T. Blazey, M. E. Raichle, Lag threads organize the brain's intrinsic activity. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E7307 (2015).
61. A. Mitra, A. Z. Snyder, C. D. Hacker, M. E. Raichle, Lag structure in resting-state fMRI. *J. Neurophysiol.* **111**, 2374–2391 (2014).
62. A. Mitra, M. E. Raichle, How networks communicate: Propagation patterns in spontaneous brain activity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150546 (2016).
63. S. A. Nastase, V. Gazzola, U. Hasson, C. Keysers, Measuring shared responses across subjects using intersubject correlation. *Soc. Cogn. Affect. Neurosci.* **14**, 667–685 (2019).
64. E. Simony *et al.*, Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* **7**, 12141 (2016).
65. M. Catani, M. Mesulam, The arcuate fasciculus and the disconnection theme in language and aphasia: History and current state. *Cortex* **44**, 953–961 (2008).
66. U. Hasson, S. A. Nastase, A. Goldstein, Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
67. S. Dehaene, L. Cohen, Cultural recycling of cortical maps. *Neuron* **56**, 384–398 (2007).
68. O. Barak, Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* **46**, 1–6 (2017).
69. M. Rigotti, D. B. D. Rubin, X. J. Wang, S. Fusi, Internal representation of task rules by recurrent dynamics: The importance of the diversity of neural responses. *Front. Comput. Neurosci.* **4**, 24 (2010).
70. I. Yamada *et al.*, Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.4, pp. 23–30.
71. L. Thiede, R. Zimmermann, Easyesn: A library for recurrent neural networks using echo state networks. (2017). https://pypi.org/project/easyesn/. Accessed 27 April 2021.
72. S. A. Nastase *et al.*, The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Sci. Data* **8**, 250 (2021).
73. B. R. Payne, C.-L. Lee, K. D. Federmeier, Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology* **52**, 1456–1469 (2015).
74. C. Pallier, A.-D.A. Devauchelle, S. Dehaene, Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 2522–2527 (2011).
75. R. Levy, Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).