PNAS RESEARCH ARTICLE | COMPUTER SCIENCES
PSYCHOLOGICAL AND COGNITIVE SCIENCES

Quantifying the narrative flow of imagined versus autobiographical stories

Maarten Sap[a,b,c,1,2], Anna Jafarpour[d,1,2], Yejin Choi[b,e], Noah A. Smith[b,e], James W. Pennebaker[f], and Eric Horvitz[b,c,1]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received July 7, 2022; accepted August 30, 2022

Lifelong experiences and learned knowledge lead to shared expectations about how common situations tend to unfold. Such knowledge of narrative event flow enables people to weave together a story. However, comparable computational tools to evaluate the flow of events in narratives are limited. We quantify the differences between autobiographical and imagined stories by introducing sequentiality, a measure of narrative flow of events, drawing probabilistic inferences from a cutting-edge large language model (GPT-3). Sequentiality captures the flow of a narrative by comparing the probability of a sentence with and without its preceding story context. We applied our measure to study thousands of diary-like stories, collected from crowdworkers, about either a recent remembered experience or an imagined story on the same topic. The results show that imagined stories have higher sequentiality than autobiographical stories and that the sequentiality of autobiographical stories increases when the memories are retold several months later. In pursuit of deeper understandings of how sequentiality measures the flow of narratives, we explore proportions of major and minor events in story sentences, as annotated by crowdworkers. We find that lower sequentiality is associated with higher proportions of major events. The methods and results highlight opportunities to use cutting-edge computational analyses, such as sequentiality, on large corpora of matched imagined and autobiographical stories to investigate the influences of memory and reasoning on language generation processes.

natural language processing | imagination | narrative | memory | deep neural networks

Significance

We explore the open question about differences in the narrative flow of stories generated from memory versus imagination. We introduce sequentiality, a computational measure of narrative flow of events that compares the influence of preceding sentences versus story topic on story sentences, using a cutting-edge large language model (GPT-3). Applying sequentiality to thousands of stories, we find that the narrative flows of imagined stories have greater reliance on preceding sentences than for autobiographical stories and that autobiographical narratives become more similar to imagined stories when retold several months later. Furthermore, we uncover a link between events perceived as salient and sequentiality. The methods provide a window into cognitive processes of storytelling that breaks away from traditional approaches to analyzing narratives.

When we tell a story, we weave together sets of events to form a coherent narrative (1–3). The narrative flow of those events is influenced by our recollection of experiences from episodic memory (4–6) as well as common knowledge about prototypical sequences of events, referred to as schema (1, 7–11). For example, telling an imagined story about a friend's wedding relies on common knowledge about the schema of how a wedding in their culture unfolds. In contrast, a recalled story drawn from memories about a friend's wedding involves an autobiographical recollection of episodic details about experienced events in addition to the knowledge of wedding schema (12). Furthermore, in autobiographical stories, the extent to which schema and episodic details are used in storytelling changes with time passing, as memories of experience become consolidated and schematized into more abstract, semantic, and "gist-like" versions (13–15).

A key element of narrative storytelling is referencing occurrences of salient events (16), which often deviate from prototypical schema (17). Such salient events can range from major (e.g., big plot twists) to minor (e.g., subtle details) (18), and from surprising to expected. Small-scale human studies have demonstrated that salient events often mark surprising or expected shifts in a story [e.g., of character focus, location, or circumstances (19)], that they stand out as particularly memorable to readers (20, 21), and that they can influence the emotional impact of a narrative (22). However, how salient events contribute to the narrative flow of imagined or autobiographical stories is not well understood.

We introduce a computational measure, sequentiality, to uncover how autobiographical and imagined stories differ with respect to narrative flow and occurrences of salient events. Sequentiality leverages probabilities of words and sentences in stories to determine the difference in the likelihood of story sentences conditioned on a story's topic versus conditioned on the story topic and the context given by all of the preceding sentences (Fig. 1). In this work, we draw probabilities from a cutting-edge and large-scale language model [GPT-3, with 175 billion parameters (23)], substantially scaling up our previous investigations (24) that employed a much smaller language model [GPT-1, with 115 million parameters (25)]. By using large-scale language models, sequentiality presents a characterization of narrative flow in stories that contrasts with previous measures which focused on either detecting event words from sentences (16, 26) or tracking attributes over time in stories [e.g., sentiment, emotion, categories of words, or sentence embeddings (27–29)].

Author contributions: M.S., A.J., Y.C., N.A.S., and E.H. designed research; M.S., A.J., and E.H. performed research; M.S., A.J., and E.H. contributed new reagents/analytic tools; M.S., A.J., and E.H. analyzed data; and M.S., A.J., Y.C., N.A.S., J.W.P., and E.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: maartensap@cmu.edu, annaja@uw.edu, or horvitz@microsoft.com.

[2]M.S. and A.J. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2211715119/-/DCSupplemental.

Published November 2, 2022.

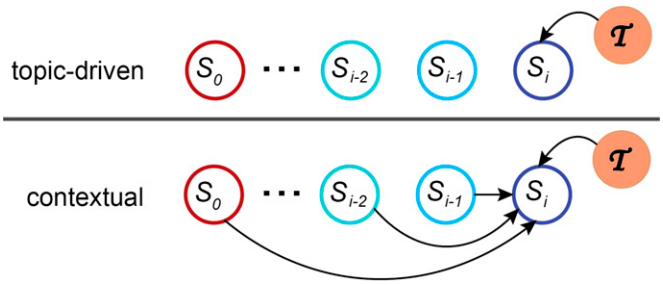

Fig. 1. Graphical models depicting the two components of sequentiality. Sequentiality reflects the probabilistic relationship among consecutive sentences ($s_0, s_1, \ldots, s_i$) in a story about a topic $\mathcal{T}$, and is computed as the difference between the log-likelihood of a sentence conditioned only on the story topic (i.e., topic-driven model; *Top*) and the log-likelihood of that sentence conditioned on both the story topic and all preceding sentences (i.e., contextual model; *Bottom*). The log-likelihood of a sentence given a topic or topic and prior sentences is provided by the GPT-3 neural language model.

We studied sequentiality and salient events in a set of 7,000 diary-like short stories about memorable life experiences, to analyze differences in narrative flow of imagined or autobiographical stories. Collected through crowdsourcing and made available in the HIPPOCORPUS dataset (24), these stories were either written about an autobiographical personal experience, recalled shortly after it happened and retold several months later, or about an imagined experience on the same topic. We extended a subset of 240 HIPPOCORPUS stories to additionally include sentence-level human annotation of event saliency. We applied sequentiality to these stories to analyze narrative flow difference in autobiographical and imagined stories, and to compare the sequentiality of sentences with various levels of event saliency. To supplement sequentiality, we also employed coarser-grained metrics that count the proportions of realis event terms—references to factual, concrete, nonhypothesized events, as well as count words in the Linguistic Inquiry Word Count (LIWC) (30) and concreteness (31) lexicons to further examine the differences in stories and event types.

We hypothesized that autobiographical and imagined stories would differ in sequentiality and event distributions, specifically, that imagined stories would have higher sequentiality, since they are more likely to follow commonly expected schema (5, 32). On the other hand, we hypothesized that autobiographical stories would have lower sequentiality but a higher number of annotated salient events, based on the intuition that those stories likely contain more-specific details drawn directly from episodic memory (5, 33) and that memorable details of a specific experience are more likely to diverge from the expected flow of the narrative (20). We also expected to find an increase in sequentiality for stories that are retold after a period of time versus freshly recalled memories, due to the consolidation and narrativization of memories over time (14).

Sequentiality for Analyzing Narrative Flow in Stories

Sequentiality provides a measure of narrative flow based on probabilities of story sentences given by large language models (LLMs). We apply the measure to identify differences in the sequencing of ideas in recalled versus imagined stories. One might expect that imagined stories composed in real time would tend to be described by a contextual model where a next sentence depends greatly on the prior sentences, with a sequencing influenced by commonly understood schemas (5, 32). In contrast, generating an autobiographical story may rely less on such an incremental sequencing and prototypical schemas (20) and be better explained by a process of organizing and building a narrative from a set of events encoded in episodic memory (5, 33).

The sequentiality metric compares, for all sentences of a story, the differences in likelihood for each sentence as predicted by a contextual sequencing model versus as predicted by a topic-driven model where each sentence is conditioned only on the topic. That is, given sentences from a story written about a topic $\mathcal{T}$, sequentiality compares the likelihood of each story sentence under two generative models, illustrated in Fig. 1. The contextual generative model assumes that each sentence is generated conditioned on the story topic as well as all of its preceding sentences. The topic-driven generative model assumes that every generated sentence is conditioned only on the story topic. As such, higher values of sequentiality for sentences suggests that the sentences follow the common expectations given the context of the evolving story and topic, whereas lower values suggest that sentences deviate more from expectation, given the preceding sequence of sentences. Here, we first briefly introduce LLMs, then formally define sequentiality, and, finally, discuss word-based narrative measures that we also use in our experiments.

Large-Scale Language Models. LLMs are a new family of language models (LMs) represented as large-scale neural networks, which have rapidly come to serve as the foundation of most current natural language processing systems (34). Formally, a language model is a statistical model that estimates the likelihood or probability of sequences of words, that is, one or more sentences. We denote this likelihood as $p_{\text{LM}}(s_{0:i})$, where $s_{0:i} = \{s_0, \ldots, s_i\}$ are consecutive sentences. LLMs are trained to estimate the likelihoods of sentences using massive amounts of text. For example, the model we use in our experiments (GPT-3) (23) is a 175-billion-parameter neural LM trained on over 45 TB of text data (e.g., books, news articles, Wikipedia pages). Through training on such large amounts of text, LLMs also learn an estimate of the general ordering or expected narrative flow of events and sentences in stories (35, 36).

Formalization. Sequentiality $c(s_i, h)$ is measured for each sentence s_i of a story about topic $\mathcal{T}$ for a number h of preceding sentences (the history under consideration, $s_{i-h:i-1}$). The $c(s_i, h)$ is computed for each sentence s_i, as the difference in the negative log-likelihood (NLL) of the sentence, as computed by the contextual and topic-driven models. This requires computing the likelihood of each sentence, conditioned on h prior sentences, per the history $s_{i-h:i-1}$ under consideration, in addition to the words in the story topic $\mathcal{T}$ and, separately, computing the likelihood of the same sentences when each is conditioned only on the story topic $\mathcal{T}$,

$$c(s_i, h) = -\frac{1}{|s_i|}[\log \underbrace{p_{\text{LM}}(s_i \mid \mathcal{T})}_{\text{topic-driven}} - \log \underbrace{p_{\text{LM}}(s_i \mid \mathcal{T}, s_{i-h:i-1})}_{\text{contextual}}],$$

where we obtain the likelihood of sentences p_{LM} from LLMs, and normalize the likelihoods by sentence length $|s_i|$ to account for sentence length variation. We then define the overall sequentiality of the entire story as the average sequentiality of its sentences.

In our analyses, we examined the average sequentiality per story for history sizes ranging from one to nine preceding sentences ($h = 1, \ldots, 9$) to the full preceding history ($h =$ full). We use the story summaries written by the storytellers as the topic $\mathcal{T}$. We compare sequentiality to the topic-driven likelihood of sentences, computed by conditioning the sentences of stories only on the topic; we report the NLL of sentences,

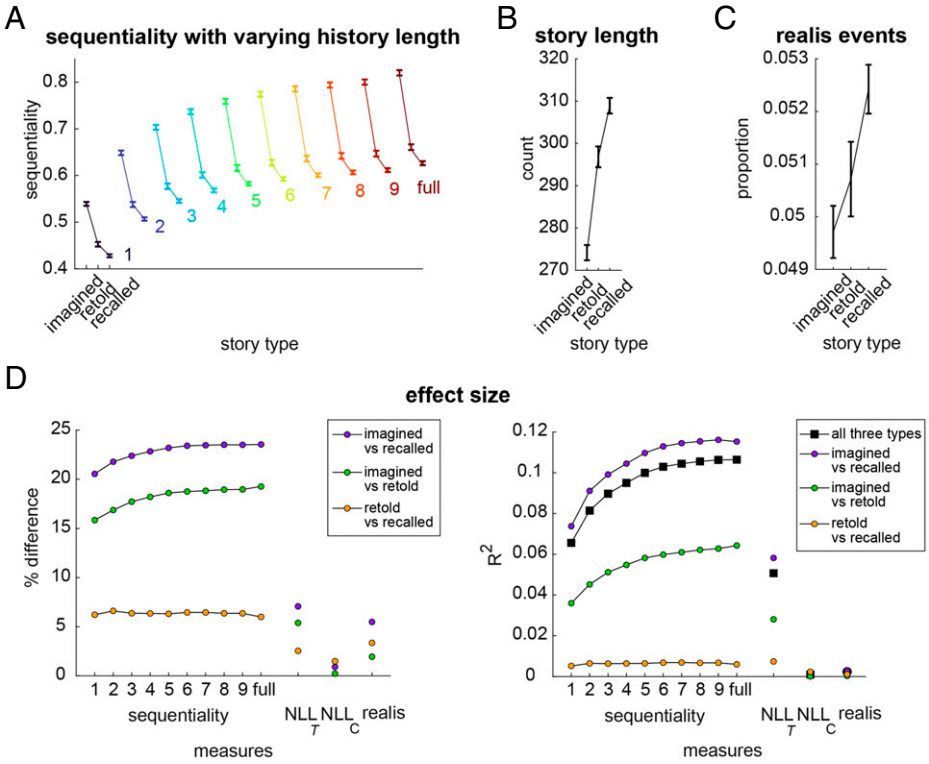

Fig. 2. Differences in sequentiality in recalled, retold, and imagined stories. (*A*) Mean sequentiality of stories with varying history lengths ($h = 1$ to $h =$ full story length) are different across the story types. Imagined stories have higher average sequentiality than autobiographical stories, and retold stories have more sequentiality than recalled stories. (*B*) Stories about imagined events are shorter than autobiographical stories. (*C*) Proportion of realis events is higher in autobiographical stories than in imagined stories. (*D*) Effect sizes: Percentage difference in parameter estimates (*Left*) and R^2 (*Right*), reflecting the magnitude of difference in sequentiality, the total number of words (story length), the topic-driven and contextual likelihoods of sentences ($\text{NLL}_{\mathcal{T}}$ and $\text{NLL}_{\mathcal{C}}$), and the proportion of realis across story types.

$\text{NLL}_{\mathcal{T}} = -1/|s_i| \log p_{\text{LM}}(s_i \mid \mathcal{T})$. We also compare to the fully contextual NLL: $\text{NLL}_{\mathcal{C}} = -1/|s_i| \log p_{\text{LM}}(s_i \mid \mathcal{T}, s_{0:i-1})$.

Lexicon-Centric Measures. In addition to sequentiality, we examined the events in narratives and employed several word-based metrics to analyze narratives. The latter lexicon-centric measures include counts of the prevalence of realis event words, that is, nonhypothetical references to concrete events that took place (e.g., "she tripped") in contrast to hypothesized events (e.g., "she feared tripping," but she did not trip). To find those words, we used an automated tagger trained on an annotated corpus of realis terms (16). We also noted average numbers of words in stories falling in psychologically related categories using the LIWC (30) lexicon, and measured the average concreteness level of words using a concreteness lexicon (31). To ensure the validity of the concepts measured by these lexicon-based measures (37), we show the most frequent words in each lexicon category along with our results.

Results

Analysis of Hippocorpus Stories. We determined the difference in sequentiality across the three story types (recalled, retold, and imagined stories), using a factorial linear regression with the story type as the grouping factor and the story length. We included the story length because recalled stories are longer than retold stories ($p = 0.001$), and retold stories are longer than imagined stories ($p < 0.001$; Fig. 2*C*). We report the R^2, which quantifies the proportion of variance in the data that is explained by the group difference, the effect size, and the p-values after correction for multiple comparisons using the Bonferroni method.

Imagined stories flow in a more expected manner than autobiographical stories. The comparisons between the sequentiality across story types ($n = 6{,}854$ stories on $n = 2{,}788$ unique topics) show that imagined stories have higher sequentiality than autobiographical memories ($p < 0.001$ for the main effect of the story type on all sequentiality history lengths; see Fig. 2 for the effect sizes). The pairwise comparisons demonstrate that imagined stories have higher sequentiality than both retold ($p < 0.001$) and recalled ($p < 0.001$) stories. While there were no differences in contextual likelihood ($\text{NLL}_{\mathcal{C}}$) between story types, we observe lower topic-driven likelihood (i.e., higher $\text{NLL}_{\mathcal{T}}$) for sentences of imagined stories versus autobiographical stories. This suggests that the sentences of imagined stories, on average, have weaker links to the topic than sentences of autobiographical stories, despite both types of sentences having strong links to the preceding sentences. However, in general, sequentiality (with increasing history size) has much larger effect sizes and R^2 compared to the likelihood or realis metrics (Fig. 2*D*), which shows that sequentiality is a better measure for capturing differences in the narratives of imagined and autobiographical stories.

Retold autobiographical stories have higher sequentiality than fresh recollections. In comparison to freshly recalled stories, stories retold after several months have higher sequentiality ($p < 0.001$), are shorter ($p < 0.001$), and contain fewer realis events ($p < 0.001$; Fig. 2). This finding demonstrates systematic shifts in the narratives of autobiographical stories with time, posing questions and framing future research on the consolidation of memories and influences of such processes on recollection. We found that participants' assessments of the frequency of recalling or retelling autobiographical stories is not associated with sequentiality but that sequentiality is negatively correlated

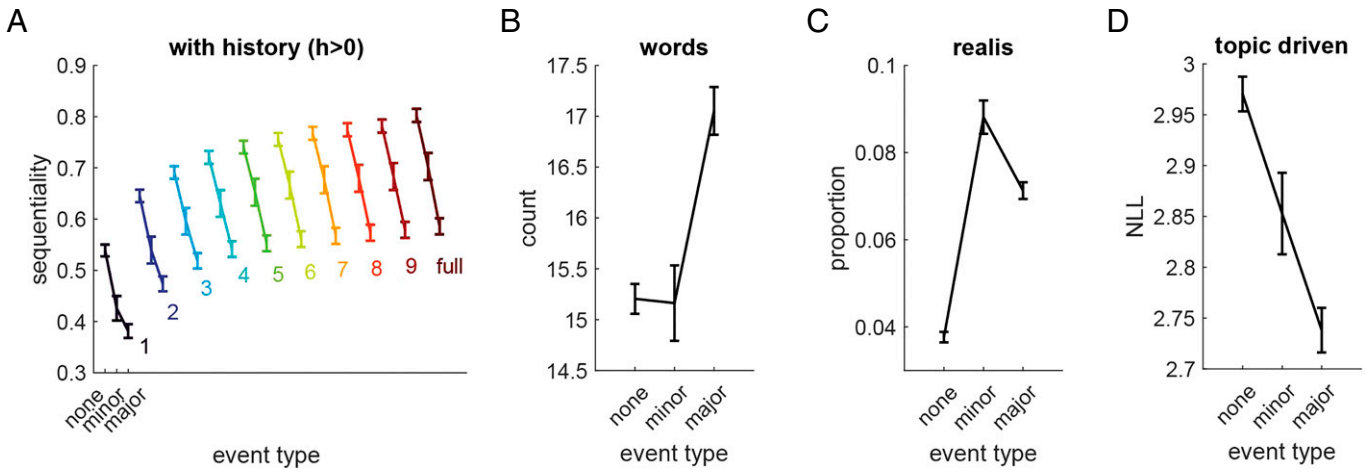

Fig. 3. Sequentiality of sentences relative to event annotations. (*A*) The average sequentiality, with varying history, is grouped by the event type. The sentences with no event (none) follow the narrative flow of the story topics more than sentences with major (all sequentiality history length) or minor events do (with sequentiality history of one sentence). Sequentialities of minor and major events are not different. (*B*) The sentences with no event are shorter than sentences with major events. (*C*) The realis in sentences with major or minor events is higher than in sentences with no event. (*D*) The NLL in sentences with major or minor events is lower than in sentences with no event. Error bars show SEM.

with the number of realis events in stories ($r = -0.08$, $p < 0.001$).

Autobiographical stories contain more realis events and concrete and time-and-space words than imagined stories. We found that the proportion of realis events is higher in recalled autobiographical stories than in imagined stories ($p = 0.001$; Fig. 2*B*), but did not differ when comparing recalled and retold ($p > 0.1$) or retold and imagined ($p > 0.1$) stories. The proportion of concrete words, measured with LIWC and concreteness lexicons (31, 38), is different across story types ($p < 0.001$; *SI Appendix*, Table S1), with fewer concrete words being used in imagined versus autobiographical stories (recall: $p < 0.001$; retold: $p < 0.001$). The proportion of concrete words is not different between recalled and retold stories ($p > 0.1$). Additionally, we found that recalled and retold stories contain greater proportions of words related to cognitive processes, time, space, and motion ($p < 0.001$; *SI Appendix*, Table S1).

Event-Annotated Subset. Next, we review the differences in the proportion of salient events in a subset of the HIPPOCORPUS that consists of 240 stories on 80 different topics across the three story types. Each story sentence was annotated by eight crowdworkers for whether a sentence expressed a major or minor event, and whether the identified event was expected versus surprising. To control for the variability in schematic knowledge and subjective understanding of what constitutes a major or minor event, the same groups of eight people annotated sentences from the three stories (imagined, recalled, retold) on each topic. We summarized the annotations based on majority voting and evaluated the difference in the proportion of major and minor events in the stories across the three story types using ANOVA including consideration of sentence length (sentences with major events are significantly longer than those with no events or with minor events; $p < 0.001$; Fig. 3*B*). Then we studied the relationships among event annotation and sequentiality, LIWC, and concreteness lexicons at the sentence level.

Autobiographical stories contain more salient events than imagined stories. We observed a main effect for story type on minor events and expected salient events, but not on major events or surprising events (Fig. 4). Specifically, higher proportions of sentences in recalled and retold stories were annotated as minor events ($p = 0.007$) and expected events ($p = 0.025$) as compared to events in imagined stories. We found no significant difference in the number of minor, major, expected, or surprising events ($p > 0.1$) between recalled stories and their retold versions.

Sentences with salient events have lower sequentiality. We examined the effect of event type (major, minor, or no event) on the sequentiality of sentences, similar to how we analyzed the effect on story types. Sequentiality with any history length shows a significant main effect of event type ($p < 0.001$; Fig. 3*A*). The sentences marked as containing major events have lower sequentiality than those with no events ($p < 0.001$, all history lengths; no difference with the minor events, $p > 0.1$), whereas, sentences with minor events have lower sequentiality than sentences with no events ($p < 0.05$) only when the sequentiality is measured considering the previous sentence ($h = 1$) but not with longer history ($h > 1$, $p > 0.1$). The results provide evidence that major events have more global influence in a story than minor events.

Sentences with salient events have a higher proportion of realis event terms and concrete, present-related, and space-related words. We found a higher proportion of realis event terms in sentences with minor events than in those with a major ($p < 0.001$) or no events ($p < 0.001$; Fig. 3*C*). Using the LIWC and concreteness lexicons, we generally observe more differences between sentences with no event and those with a salient event, compared to between sentences with a minor vs. major event (*SI Appendix*, Table S2). Notably, in addition to lower proportions of concrete and time-and-space words, we see higher proportions of words related to cognitive and affective processes in sentence with no events.

Sentences with surprising events have lower sequentiality than those with expected major events. We found that major events are often annotated as surprising (72%) rather than expected (28%), whereas all minor events are annotated as expected. Sentences annotated as describing major events have a lower sequentiality when they are noted to be surprising versus expected ($p < 0.001$). Sequentiality is also lower for expected major events compared to expected minor events ($p < 0.001$; the difference increased with increasing history length). In general, we found that sequentiality of sentences is not different for surprising and expected sentences ($p > 0.05$; the difference decreased with increasing history length; for $h = 1$, uncorrected $p = 0.014$), suggesting that sequentiality captures more than the event expectancy.

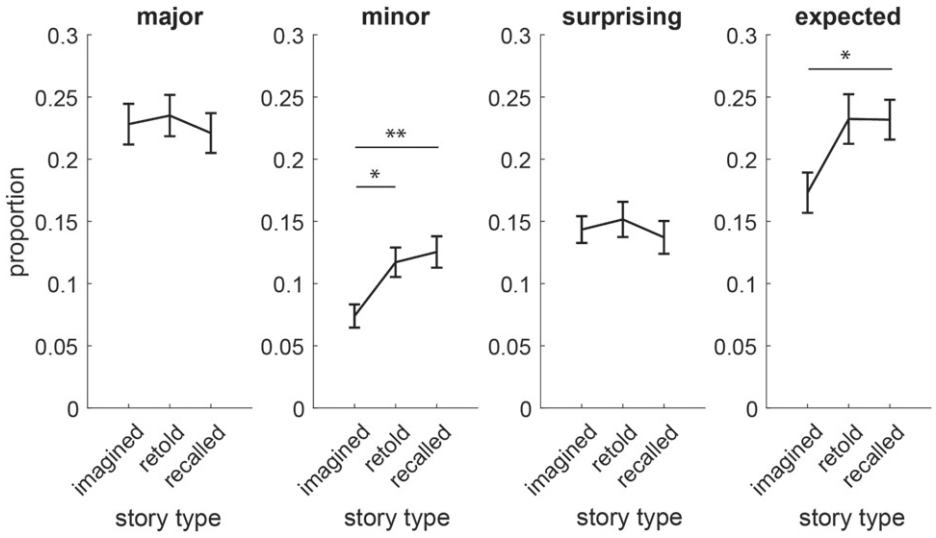

Fig. 4. Proportions of salient event annotations across the stories. Graphs of the mean and SEM of the proportion of events annotated as salient (from left: major, minor, surprising, and expected events) in the imagined, recalled, and retold stories. (*$p < 0.05$, **$p < 0.01$).

Discussion

We introduced sequentiality as a computational measure of narrative flow of events instantiated by large-scale neural language models. We used the measure to probe hypotheses about the generative processes of constructing experienced versus imagined stories, paired via a matched topical description starting point. Sequentiality measures the extent to which story events and sentences flow from their preceding context and overall story topic versus from only the story topic (Fig. 1), using likelihoods given by the GPT-3 large-scale language model (23). As such, sequentiality can be considered a proxy for quantifying how much a story follows the expected or common narrative flow for a specific story topic (schematic knowledge) versus is grounded in experiential details (episodic memory).

We used sequentiality to study differences in narrative flow across 1) recalled stories based on fresh autobiographical experiences, 2) retold stories about those same autobiographical experiences after 3 mo to 6 mo, and 3) imagined stories matched to the topics of the autobiographical stories. With sequentiality and word-based metrics such as the count of realis event terms (that refer to concrete, nonhypothesized event occurrences) and LIWC and concreteness lexicon scores, we observed differences in episodic details and differing reliance on schematic knowledge for constructing narratives. Based on sequentiality differences, imagined stories have greater alignment with expected schematic flow of events than autobiographical stories. Autobiographical stories contain more minor detailed events than imagined stories (Fig. 4), and they tend to have higher proportions of concrete words as well as words related to time and space (*SI Appendix*, Table S1). Below, we discuss implications of our findings for analyzing narrative flow of events using large-scale neural LMs, as well as for understanding cognitive processes of storytelling with computational methods.

Using Sequentiality to Quantify Narrative Flow. Sequentiality quantifies the extent to which the flow of events follows expected schema, using large-scale neural language models. This is a departure from previous measures of narrative flow, which have predominantly approached the task by examining word usage, such as the rates of emotion-related words over time. In prior work, researchers have argued that emotional flow plays a role in the persuasiveness of stories (22), an approach which was later operationalized through word counting of emotion words in books (27) and consumer reviews (39). In addition to emotion words, recent work computed the progression of the rate of function words and words related to cognitive processes, to study narrative progression and their relationship to story quality (28). Beyond studying word frequencies, a recent study (29) employed high-dimensional word vectors, to compute the speed and complexity of stories. In contrast to previous work which analyzed narratives through surface-level features, sequentiality leverages a story's topic and the language modeling capabilities of large-scale neural LMs to infer the predictability of words. Sequentiality does not rely on specific word categories or high-dimensional word vectors. Sequentiality was initially used to measure the linearity of sentences, in a preliminary investigation (24) where we used a much smaller neural LM [GPT-1 (25)] than the one used in this study.

Conceptually, the sequentiality generative model provides a lens on how sentences and events are produced or read, adding to several models of sentence and event processing from cognitive science. Sequentiality relates to word-level surprisal theory (40, 41), which posits that humans form expectations of which word should come next in text, before observing it. Contextual generative models can formalize those expectations (e.g., Fig. 1, *Bottom*), and neural LMs can approximate these human expectations about words given sufficient context (42).

However, surprisal theory does not account for the variation in noncontextual likelihood of events depending on the story topic, which may play a role in how humans form expectations. For example, a story about driving on a highway for 30 mi might have fewer expected events than one about a birthday party, which has opportunities for details on whose birthday it was, where it took place, who attended, how the cake tasted, etc. We account for this variation by conditioning both the topic-driven and contextual models on the story topic. Although we find no differences in contextual likelihood ($NLL_{\mathcal{C}}$) and only small differences in topic-driven likelihoods ($NLL_{\mathcal{T}}$), the largest difference across story types is measured as the ratio of contextual and topic-driven likelihoods using sequentiality (Fig. 2*D*). Corroborating this need for comparing likelihoods, recent work has shown the usefulness of comparing contextual and noncontextual event likelihoods in visual event segmentation tasks (21).

Sequentiality is built on the assumption that large-scale neural language models encode knowledge about the commonly expected narrative flow of events. Previous work suggests that this is a valid assumption, since LLMs can determine the correct ordering of sentences in text (25, 43) and can be used to generate expected schemas for events (35). However, the extent to which LLMs learn the common flow of events is influenced by the knowledge contained in their training data (44). Specifically, the culture and identities of the authors of training data can influence the schema that are deemed likely by the model; a language model trained exclusively on British text only will likely learn British-specific schema (e.g., tea time) that other models might not encode. However, our findings with sequentiality remain similar when using language models trained on other datasets (24), such as OpenAI-GPT [trained on 5 GB of English fiction (25)] and GPT-2 [trained on 40 GB of news-like English text (43)], suggesting this may not be a substantial issue.

Cognitive Processes of Recalling Versus Imagining. The results reveal differences in the cognitive processes of how people form narratives grounded in their own experiences versus from their imagination, and the differential role of salient events in both types of storytelling. Although imagination and remembering may engage similar mental processes (45), and imagination could leverage one's own life experiences (46), we found series of systematic differences between imagined and autobiographical stories. In all stories, storytellers appear to combine schematic knowledge with references to major events. We found that major events tend to be relayed in surprising sentences that tend to deviate from expectation, per likelihoods provided by the neural language model. These sentences are associated with the lowest sequentiality (Figs. 3 and 4), and they are often about personal concerns and core drives and needs (*SI Appendix*, Table S2). For example, in the recalled story on "A warm summer morning with a hummingbird. How I had a communal moment with nature by misting a hummingbird with a garden hose," the major event is that "At first, I thought he [the hummingbird] was just doing his early morning pollen rituals, but to my surprise he wanted water." In an imagined story on the same topic, the major event is that "[animal started to come to the garden.] Mostly squirrels at first and a few deer, and one tiny hummingbird." Similarly, in the recalled story, the major event is that "I saw a hummingbird at the corner of my eye."

A significant difference between the autobiographical and imagined stories is in the proportion of minor events, as identified through human annotations (Fig. 4). The minor events tend to be nonhypothesized, concrete details of the stories that are noted as expected but typically not part of the general schema of the story topic. The minor events have local saliency and can be identified only with computation of sequentiality with a one-sentence history. These events often contain words on biological processes and social references. As an example, a minor event in a recalled story on the same topic as the example above is that "I was feeling kind of low due to not seeing many of my friends anymore due to everyone being busy with their schedule, and work being a little slow was also on my mind," and, in an imagined story, was that "For the first few weeks I got nothing and no activity, then about a month ago animals came."

We found that sentences annotated as describing salient events tend to have more concrete words, first-person references, social words, and words related to cognitive processes, biological processes, core drives and needs, and relativity to time, space, and motion. Only a subset of these observations, including the change in time, character, and space, has been previously reported in studies on detection of salient events to mark an event boundary (2). We also observed that the length of stories showed small differences among the story types. This observation on length is congruent with the understanding that the stories that rely largely on commonly expected schema are generally shorter (47, 48).

We found that the proportions of salient events (major and minor) are similar in stories about freshly recalled memories and about memories retold after 3 mo to 6 mo (Fig. 4). The retold stories have higher sequentiality and are shorter than the initial recall of stories (Fig. 3). The self-reported frequency of revisiting and retelling autobiographical stories does not appear to influence the sequentiality of the stories. The retold stories that were noted as more frequently revisited memories were found to contain fewer realis events, which may reflect processes of abstraction. The sequentiality measure provides a means of quantifying the observation that, with passing time and memory consolidation, retelling autobiographical memories relies less on recall from episodic memory, and instead increasingly invokes common semantic knowledge of schema (1, 10, 14), especially since certain events may be forgotten (49).

Open Research Directions. The methods and results presented show promise as tools for exploring processes of memory, reasoning, and imagination employed to generate narratives. The methods also hold the opportunity to help with building deeper understandings of influences of common schema and personal experiences on the stories that people tell. From a computational perspective, we see rich opportunities ahead for harnessing large-scale neural models to explore narrative theories, including consideration and comparative study of different generative models (50). From a cognitive perspective, directions include pursuing answers to standing questions about the contributions of memory and reasoning to the stories that people generate about experienced and imagined events, and how memories—and the autobiographical stories that flow from them—evolve over time since events are experienced. From a cultural perspective, the methods can provide the opportunity to study differences across communities and cultures of the nature and influences of common schema and personal experiences on stories. Opportunities for study include seeking insights about the influences and interpretations of world events over time on fiction and nonfiction narrativizations (51, 52). Other research directions include applying the results, methods, and measures in studies of narrativizations with different motivations (53) such as recall, storytelling, persuasion, lie detection, false confessions, recovered memories, and the propagation and effects of misinformation.

Materials and Methods

Building HIPPOCORPUS. In our analyses, we make use of our previously collected corpus of autobiographical, imagined, and retold stories [HIPPOCORPUS (24)]. This corpus contains 6,854 stories collected from crowdworkers in three stages (depicted in *SI Appendix*, Fig. S1). In the recalled stage, workers write a short diary-like story and a short summary. Then, in the imagined stage, workers are given a summary and asked to write a short diary-like story. Finally, in the retold stage, workers from the first stage are given their original summary and asked to retell their story, after 3 mo to 6 mo have passed. For both the recalled and retold tasks, we collect, from workers, the time elapsed since they experienced the event (TIMESINCEEVENT, in weeks or months), as well as the frequency at which they thought or talked about the event (FREQOFRECALL, on a five-point Likert scale of "never" to "constantly"). This study was undertaken following approval from the Institutional Review Board (IRB) at Microsoft Research. For more details, see our preliminary work (24) and *SI Appendix*.

Collecting Event Annotations. We additionally collected sentence-level event annotations for a subset of the HIPPOCORPUS stories. We randomly selected 80 topics and their associated recalled–imagined–retold stories ($n = 240$ stories). Since people's individual perceptions of what constitutes an expected, surprising,

major, or minor event could differ depending on their experiences, background, or culture, we made sure all stories about the same topic were annotated by the same worker. We collected event annotations from eight crowdworkers per set of three stories. This effort was also undertaken with IRB approval, and all crowdworkers had informed consent when they chose to participate.

Participants read each of the sentences in each of the three stories, one sentence at a time, and indicated whether the story sentences mark a start of a new event. Specifically, annotators marked whether a sentence represented a new event that is minor or major and whether the events are expected or unexpected. See *SI Appendix* for further details.

Extracting Sequentiality, Realis Events, and Lexicon Counts. To compute the sequentiality of each story sentence, we first split each story in the HIPPOCORPUS into sentences, using a version of the Natural Language Toolkit (NLTK) sentence tokenizer adapted to avoid splitting sentences into one-word sentences (54). We then used the OpenAI Application Programming Interface (API) to obtain the likelihoods under GPT-3 of each sentence conditioned on the story topic and various history sizes. Specifically, we compute the log-likelihood of a sentence $p_{\text{GPT-3}}(s_i)$ by summing the word-level log-probabilities yielded by the API for the sentence at hand. We can then compute sequentiality for each history size.

For computing the proportions of realis event terms, we use a realis term tagger from our preliminary investigations (24). This tagger is a Bidirectional Encoder Representations from Transformers (BERT) (55) model trained on a realis annotated corpus of literary fiction (16), which achieves F1 accuracy scores of 83.7% and 75.8%, on the validation and test sets, respectively.

We used the LIWC 2015 software for counting the proportion of words that belong to specific LIWC categories (38). For the concreteness lexicon (31), we averaged the concreteness lexicon of each story by matching words in the story with words in the lexicon.

Data Analysis. For each story, we averaged the sequentiality of all sentences and had one representative value for each of the sequentialities with history length of one sentence to full story. We also took the averaged proportion of major or minor events, the averaged proportion of realis events, the total number of words (story length), and the averaged NLL ($\text{NLL}_{\mathcal{C}}$, $\text{NLL}_{\mathcal{T}}$) per story. We applied a factorial linear regression on each of the parameters, to identify the differences between story types. We either included three factors for three story types (imagined, retold, recalled) or included two factors for pairwise comparisons.

We similarly used a factorial linear regression to evaluate the characteristics of sentences with various event types (major, minor, or no events). A sentence was accepted to be a minor or major event if the majority of the annotators marked the sentence as such. We also evaluated the proportion of events that were expected or surprising by the majority of the annotators. This analysis was done at the sentence level with 9,412 major, 6,835 minor, and 17,477 no event annotation. We use Bonferroni correction to adjust the significance threshold for multiple comparisons. All reported *P* values are Bonferroni corrected.

Data, Materials, and Software Availability. A dataset of 6,854 English diary-like short stories about recalled and imagined events, including tagging of events, has been deposited in Hippocorpus (https://msropendata.com/datasets/0a83fb6f-a759-4a17-aaa2-fbac84577318) (24).

ACKNOWLEDGMENTS. Microsoft provided an internship for M.S, support for crowdworkers, and computing resources for running the GPT-3 neural language model on HIPPOCORPUS. A.J. was supported by NIH Brain Initiative Grant K99MH120048. We are indebted to Paul Koch at the Office of the Chief Scientific Officer for for running the GPT-3 computation. We thank Zhilin Wang and members of the Buffalo Lab at the University of Washington for valuable discussions.

Author affiliations: [a]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA 98195; [c]Microsoft, Redmond, WA 98052; [d]Department of Physiology and Biophysics, University of Washington, Seattle, WA 98195; [e]Allen Institute for AI, Seattle, WA 98103; and [f]Department of Psychology, University of Texas, Austin, TX 78712

1. F. C. Bartlett, *Remembering: A Study in Experimental and Social Psychology* (Cambridge University Press, 1932).
2. C. A. Kurby, J. M. Zacks, Segmentation in the perception and memory of events. *Trends Cogn. Sci.* **12**, 72–79 (2008).
3. J. B. Black, H. Bern, Causal coherence and memory for events in narratives. *J. Verbal Learn. Verbal Behav.* **20**, 267–275 (1981).
4. M. A. Conway, A. F. Collins, S. E. Gathercole, S. J. Anderson, Recollections of true and false autobiographical memories. *J. Exp. Psychol. Gen.* **125**, 69–95 (1996).
5. M. A. Conway, C. W. Pleydell-Pearce, S. E. Whitecross, H. Sharpe, Neuropsychological correlates of memory for experienced and imagined events. *Neuropsychologia* **41**, 334–340 (2003).
6. E. Tulving, Episodic and semantic memory. *Organ. Mem* **1**, 381–403 (1972).
7. R. C. Schank, R. P. Abelson, *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures* (Lawrence Erlbaum, 1977).
8. W. Kintsch, The role of knowledge in discourse comprehension: A construction-integration model. *Psychol. Rev.* **95**, 163–182 (1988).
9. A. C. Graesser, S. P. Robertson, P. A. Anderson, Incorporating inferences in narrative representations: A study of how and why. *Cognit. Psychol.* **13**, 1–26 (1981).
10. G. H. Bower, J. B. Black, T. J. Turner, Scripts in memory for text. *Cognit. Psychol.* **11**, 177–220 (1979).
11. I. E. Hyman Jr., E. F. Loftus, Errors in autobiographical memory. *Clin. Psychol. Rev.* **18**, 933–947 (1998).
12. A. Gilboa, R. S. Rosenbaum, A. Mendelsohn, Autobiographical memory: From experiences to brain representations. *Neuropsychologia* **110**, 1–6 (2018).
13. M. T. R. van Kesteren, D. J. Ruiter, G. Fernández, R. N. Henson, How schema and novelty augment memory formation. *Trends Neurosci.* **35**, 211–219 (2012).
14. A. Smorti, C. Fioretti, Why narrating changes memory: A contribution to an integrative model of memory and narrative processes. *Integr. Psychol. Behav. Sci.* **50**, 296–319 (2016).
15. D. Clewett, S. DuBrow, L. Davachi, Transcending time in the brain: How event memories are constructed from experience. *Hippocampus* **29**, 162–183 (2019).
16. M. Sims, J. H. Park, D. Bamman, *Literary Event Detection* (Association for Computational Linguistics, 2019).
17. R. A. Zwaan, G. A. Radvansky, Situation models in language comprehension and memory. *Psychol. Bull.* **123**, 162–185 (1998).
18. A. Jafarpour, S. Griffin, J. J. Lin, R. T. Knight, Medial orbitofrontal cortex, dorsolateral prefrontal cortex, and hippocampus differentially represent the event saliency. *J. Cogn. Neurosci.* **31**, 874–884 (2019).
19. J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, Event perception: A mind-brain perspective. *Psychol. Bull.* **133**, 273–293 (2007).
20. R. Reichardt, B. Polner, P. Simor, Novelty manipulations, memory performance, and predictive coding: The role of unexpectedness. *Front. Hum. Neurosci.* **14**, 152 (2020).
21. N. T. Franklin, K. A. Norman, C. Ranganath, J. M. Zacks, S. J. Gershman, Structured Event Memory: A neuro-symbolic model of event cognition. *Psychol. Rev.* **127**, 327–361 (2020).
22. R. L. Nabi, M. C. Green, The role of a narrative's emotional flow in promoting persuasive outcomes. *Media Psychol.* **18**, 137–162 (2015).
23. T. B. Brown *et al.*, Language models are few-shot learners. *Adv. Neural Inf. Process Syst.* **33**, 1877–1901 (2020).
24. M. Sap, E. Horvitz, Y. Choi, N. A. Smith, J. W. Pennebaker, "Recollection versus imagination: Exploring human memory and cognition via neural language models" in *Proceedings of the Association for Computational Linguistics*, 1970–1978 (Association for Computational Linguistics, Seattle, WA, 2020).
25. A Radford, K Narasimhan, T Salimans, I Sutskever, Improving language understanding by generative pre-training (2018). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed 17 October 2022.
26. B. Li, S. Lee-Urban, G. Johnston, M. Riedl, "Story generation with crowdsourced plot graphs" in *Twenty-Seventh AAAI Conference on Artificial Intelligence* **27**, 598–604 (Association for the Advancement of Artificial Intelligence, 2013).
27. A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, P. S. Dodds, The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **5**, 31 (2016).
28. R. L. Boyd, K. G. Blackburn, J. W. Pennebaker, The narrative arc: Revealing core narrative structures through text analysis. *Sci. Adv.* **6**, eaba2196 (2020).
29. O. Toubia, J. Berger, J. Eliashberg, How quantifying the shape of stories predicts their success. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2011695118 (2021).
30. Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
31. M. Brysbaert, A. B. Warriner, V. Kuperman, Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
32. K. Michaelian, Episodic and semantic memory and imagination: The need for definitions. *Am. J. Psychol.* **131**, 99–103 (2018).
33. M. A. Greenberg, C. B. Wortman, A. A. Stone, Emotional expression and physical health: Revising traumatic memories or fostering self-regulation? *J. Pers. Soc. Psychol.* **71**, 588–602 (1996).
34. R. Bommasani *et al.*, On the opportunities and risks of foundation models. *arXiv* [Preprint] (2021). arXiv:2108.07258. https://crfm.stanford.edu/assets/report.pdf. Accessed 17 June 2022.
35. P. Laban, L. Dai, L. Bandarkar, M. A. Hearst, "Can transformer models measure coherence in text: Re-thinking the shuffle test" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Association for Computational Linguistics, 2021), pp. 1058–1064.
36. E. Clark *et al.*, "All that's 'human' is not gold: Evaluating human evaluation of generated text" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Association for Computational Linguistics, 2021), pp. 7282–7296.
37. K. Jaidka *et al.*, Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10165–10171 (2020).
38. J. W. Pennebaker, R. J. Booth, R. L. Boyd, M. E. Francis, *Linguistic Inquiry and Word Count: LIWC 2015* (Pennebaker Conglomerate, 2015).
39. T. van Laer, J. Edson Escalas, S. Ludwig, E. A. van den Hende, What happens in Vegas stays on TripAdvisor? A theory and technique to understand narrativity in consumer reviews. *J. Consum. Res.* **46**, 267–285 (2019).

40. J. Hale, "A probabilistic earley parser as a psycholinguistic model" in *NAACL-HLT* (Association for Computational Linguistics, 2001), pp. 1–8.
41. R. Levy, Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
42. A. Goldstein *et al.*, Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv* [Preprint] (2021). https://doi.org/10.1101.2020.12.02.403477. Accessed 17 June 2022.
43. A Radford, et al., Language models are unsupervised multitask learners (2019). https://pdfs.semanticscholar.org/41f9/45f59bd0d345d4e355fb72110524f6fdffdb.pdf. Accessed 17 October 2022.
44. J. Gordon, B. Van Durme, "Reporting bias and knowledge acquisition" in *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction* (Association for Computing Machinery, 2013), pp. 25–30.
45. D. L. Schacter, D. R. Addis, The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 773–786 (2007).
46. A. Grysman, J. A. Hudson, The self in autobiographical memory: Effects of self-salience on narrative content and structure. *Memory* **19**, 501–513 (2011).
47. P. W. Thorndyke, Cognitive structures in comprehension and memory of narrative discourse. *Cognit. Psychol.* **9**, 77–110 (1977).
48. W. Kintsch, E. Greene, The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Process.* **1**, 1–13 (1978).
49. L. R. Squire, Two forms of human amnesia: An analysis of forgetting. *J. Neurosci.* **1**, 635–640 (1981).
50. A. Piper, R. J. So, D. Bamman, "Narrative theory for computational narrative understanding" in *EMNLP 2021* (Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021), pp. 298–311.
51. T. Underwood, *The Invention of Historical Perspective* (Stanford University Press, 2013), pp. 55–80.
52. T. Underwood, Machine learning and human perspective. *PMLA* **135**, 92–109 (2020).
53. E. J. Marsh, Retelling is not the same as recalling: Implications for memory. *Curr. Dir. Psychol. Sci.* **16**, 16–20 (2007).
54. NLTK Project, NLTK documentation. https://www.nltk.org/api/nltk.tokenize.html. Accessed 20 May 2020.
55. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (NAACL, 2019).