Quantifying the narrative flow of imagined versus autobiographical stories

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

[a]Language Technologies Institute, Carnegie Mellon University; [b]Paul G. Allen School of Computer Science & Engineering, University of Washington; [c]Department of Physiology and Biophysics, University of Washington; [d]Microsoft; [e]Allen Institute for AI; [f]University of Texas at Austin; [*]equal contribution

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 | autobiographical memory | memory consolidation | imagination | deep neural network | pretrained language models

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 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 novel 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 175B parameters; 23), substantially scaling up our previous investigations (24) that employed a much smaller language model (GPT-1, with 115M parameters; 25). By using large-scale language models, sequentiality presents a novel characterization of narrative flow in stories, contrasting with previous measures which either focused on 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).

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 autobio-

Significance Statement

We explore the open question about differences in 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 new window into cognitive processes of storytelling that breaks away from traditional approaches to analyzing narratives.

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXX

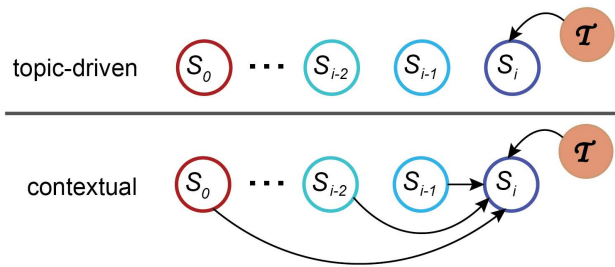PNAS | September 30, 2022 | vol. XXX | no. XX | **1–11**

Fig. 1. **Graphical models depicting the two components of sequentiality.** Sequentiality reflects the probabilistic relationship among consecutive sentences $(s_0, s_1, ..., s_i)$ 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 row) and the log-likelihood of that sentence conditioned on both the story topic and all preceding sentences (i.e., *contextual* model; bottom row). The log-likelihood of a sentence given a topic or topic and prior sentences is provided by the GPT-3 neural language model.

graphical stories. Collected through crowdsourcing and made available in the HIPPOCORPUS data set (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, non-hypothesized events, as well as count words in the 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 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 NLP systems (34). Formally, a language model is a statistical model that estimates the *likelihood* or probability of sequences of words, i.e., one or more sentences. We denote this likelihood as $p_{\text{LM}}(s_{0:i})$ where $s_{0:i} = \{s_0, ..., s_i\}$ 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 175B parameter neural LM trained on over 45TB 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, $s_{i-h:i-1}$). $c(s_i, h)$ 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
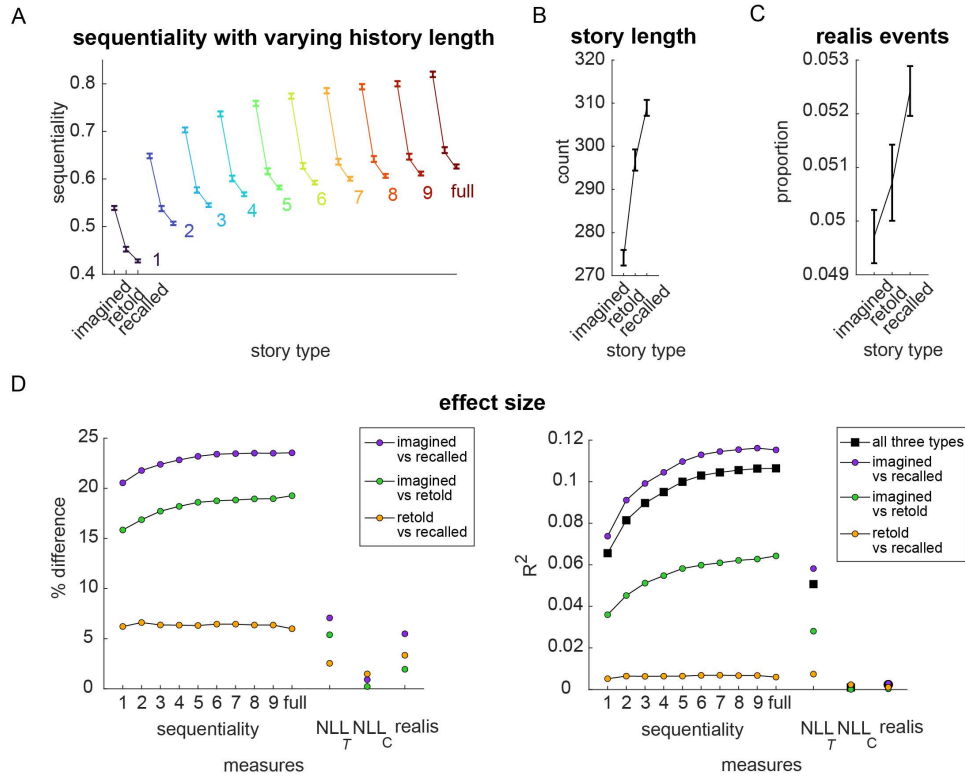
Fig. 2. Differences in sequentiality in recalled, retold, and imagined stories. (A) Mean sequentiality of stories with varying history lengths ($h = 1$ to h = full story length) are different across the story types. Imagined stories have higher average sequentiality than autobiographical stories, and retold stories 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.

sentences ($h = 1, ..., 9$) to the full preceding history ($h = \text{full}$). 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 negative log-likelihood of sentences, $\text{NLL}_\mathcal{T} = -\frac{1}{|s_i|} \log p_{\text{LM}}(s_i \mid \mathcal{T})$. We also compare to the fully contextual negative log-likelihood: $\text{NLL}_\mathcal{C} = -\frac{1}{|s_i|} \log p_{\text{LM}}(s_i \mid \mathcal{T}, s_{0:i-1})$.

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, i.e., non-hypothetical 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 Linguistic Inquiry Word Count (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. 2C). 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 ($N = 6854$ stories on $N = 2788$ 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. 2D), 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 in 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 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. 2B), 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 ($p < 0.001$; supplementary Tab. 1) 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 ($p < 0.001$; supplementary Table. 1).

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. 4B). 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. 3). Specifically, higher proportions of sentences in recalled and retold stories were annotated as minor events ($p = 0.007$) and expected events ($p = 0.025$)
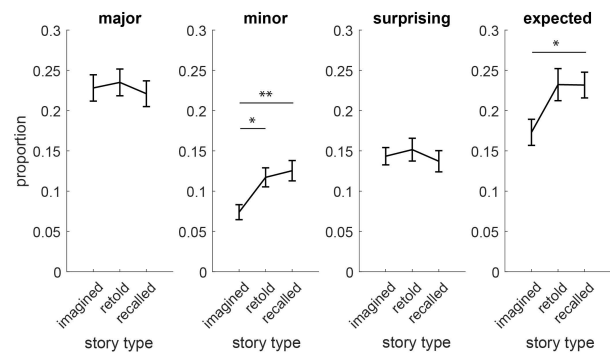

Fig. 3. Proportions of salient event annotations across the stories. Graphs of the mean and standard error of the mean of the proportion of events annotated as salient (from left: major, minor, surprising, and expected events) in the imagined, recalled, and retold stories. (* $p < 0.05$, ** $p < 0.01$)

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 show a significant main effect of event type ($p < 0.001$; Fig. 4A). 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 proportions 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. 4C). 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 (see supplementary Table 2). 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
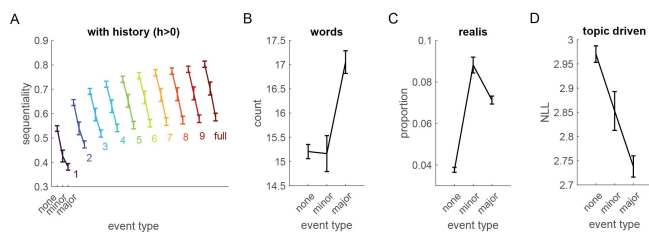
Fig. 4. Sequentiality of sentences relative to event annotations. (A) The average sequentiality, with varying history, is grouped by the event type. The sentences with no event (none) follow the narrative flow of the story topics more than sentences with major (all sequentiality history length) or minor events do (with sequentiality history of one sentence). Sequentiality 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. Error bars show standard error of the mean.

decreased with increasing history length; for h = 1, uncorrected $p = 0.014$), suggesting that sequentiality captures more than the event expectancy.

Discussion

We introduced *sequentiality* as a new 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 three to six months, 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, non-hypothesized 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. 3), and they tend to have higher proportions of concrete words as well as words related to time and space (supplementary Table 1). 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 is a novel measure to quantify 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 (40) 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 new 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* (41, 42), 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 (43).

However, surprisal theory does not account for the variation in non-contextual 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 miles 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. 2D). Corroborating this need for comparing likelihoods, recent work has shown the usefulness of comparing contextual and non-contextual event likelihoods in a 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 sentence in text (25, 44) 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 (45). 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 data sets (24), such as OpenAI-GPT (trained on 5GBs of English fiction; 25) and GPT-2 (trained on 40GBs of news-like English text; 44), 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 (46) and imagination could leverage one's own life experiences (47), 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 (Fig. 3 and Fig. 4), and they are often about personal concerns and core drives and needs (supp. Table 2). 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. 3). The minor events tend to be non-hypothesized, 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.*"

Shedding some new light on the nature of salient events, 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 (48). 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 (49, 50).

We found that the proportion of salient events (major and minor) are similar in stories about freshly recalled memories and about memories retold after 3–6 months (Fig. 3). The retold stories have higher sequentiality and are shorter than the initial recall of stories (Fig. 4). 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, instead increasingly invokes common semantic knowledge of schema (1, 10, 14), especially since certain events may be forgotten (51).

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 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 (52). 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 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 non-fiction narrativizations (53, 54). Other research directions include applying the results, methods, and measures in studies of narrativizations with different motivations (55) 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 Fig. 5). 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 re-tell their story, after 3-6 months 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 the Supplementary materials A.

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 make sure all stories about the same topic are annotated by the same worker. We collected event annotations from 8 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 if 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 if the events are *expected* or *unexpected*. See Supplementary B 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 NLTK sentence tokenizer* adapted to avoid splitting sentences into one-word sentences. We then used the OpenAI 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 BERT [56] 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 sequentiality with history length of 1 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 negative log likelihoods ($\text{NLL}_{\mathcal{C}}$, $\text{NLL}_{\mathcal{T}}$) 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 we included two factors for pairwise comparisons.

We similarly used a factorial linear regression to evaluate the characteristics of sentences with various event type (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 9412

---

* https://www.nltk.org/api/nltk.tokenize.html
† https://www.liwc.app/

major, 6835 minor, and 17477 no event annotation. We used Bonferroni correction to adjust the significance threshold for multiple comparisons. All reported p-values are Bonferroni corrected.

1. FC Bartlett, *Remembering: A study in experimental and social psychology.* (Cambridge University Press), (1932).
2. CA Kurby, JM Zacks, Segmentation in the perception and memory of events. *Trends cognitive sciences* **12**, 72–79 (2008).
3. JB Black, H Bern, Causal coherence and memory for events in narratives. *J. Verbal Learn. Verbal Behav.* **20**, 267–275 (1981).
4. MA Conway, AF Collins, SE Gathercole, SJ Anderson, Recollections of true and false autobiographical memories. *J. Exp. Psychol. Gen.* **125** (1996).
5. MA Conway, CW Pleydell-Pearce, SE Whitecross, H Sharpe, Neurophysiological 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. RC Schank, RP 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 (1988).
9. AC Graesser, SP Robertson, PA Anderson, Incorporating inferences in narrative representations: A study of how and why. *Cogn. Psychol.* **13**, 1–26 (1981).
10. GH Bower, JB Black, TJ Turner, Scripts in memory for text. *Cogn. psychology* **11**, 177–220 (1979).
11. IE Hyman Jr, EF Loftus, Errors in autobiographical memory. *Clin. psychology review* **18**, 933–947 (1998).
12. A Gilboa, RS Rosenbaum, A Mendelsohn, Autobiographical memory: From experiences to brain representations. *Neuropsychologia* **110**, 1–6 (2018).
13. MTR van Kesteren, DJ Ruiter, G Fernández, RN Henson, How schema and novelty augment memory formation. *Trends Neurosci.* **35** (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, JH Park, D Bamman, Literary event detection in *ACL.* (2019).
17. RA Zwaan, GA Radvansky, Situation models in language comprehension and memory. *Psychol. bulletin* **123**, 162 (1998).
18. A Jafarpour, S Griffin, JJ Lin, RT Knight, Medial orbitofrontal cortex, dorsolateral prefrontal cortex, and hippocampus differentially represent the event saliency. *J. cognitive neuroscience* **31**, 874–884 (2019).
19. JM Zacks, NK Speer, KM Swallow, TS Braver, JR Reynolds, Event perception: a mind-brain perspective. *Psychol. bulletin* **133**, 273 (2007).
20. R Reichardt, B Polner, P Simor, Novelty manipulations, memory performance, and predictive coding: The role of unexpectedness. *Front. human neuroscience* **14**, 152 (2020).
21. NT Franklin, KA Norman, C Ranganath, JM Zacks, SJ Gershman, Structured event memory: A neuro-symbolic model of event cognition. *Psychol. Rev.* **127**, 327 (2020).
22. RL Nabi, MC Green, The role of a narrative's emotional flow in promoting persuasive outcomes. *Media Psychol.* **18**, 137–162 (2015).
23. TB Brown, et al., Language models are few-shot learners. unpublished (2020).
24. M Sap, E Horvitz, Y Choi, NA Smith, JW Pennebaker, Recollection versus imagination: Exploring human memory and cognition via neural language models, in *Proceedings of the Association for Computational Linguistics.* (Association for Computational Linguistics, Seattle, Washington), (2020).
25. A Radford, K Narasimhan, T Salimans, I Sutskever, Improving language understanding by generative pre-training. unpublished (2018).
26. B Li, S Lee-Urban, G Johnston, M Riedl, Story generation with crowdsourced plot graphs in *Twenty-Seventh AAAI Conference on Artificial Intelligence.* (2013).
27. AJ Reagan, L Mitchell, D Kiley, CM Danforth, PS Dodds, The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **5**, 31 (2016).
28. RL Boyd, KG Blackburn, JW Pennebaker, The narrative arc: Revealing core narrative structures through text analysis. *Sci. advances* **6**, eaba2196 (2020).
29. O Toubia, J Berger, J Eliashberg, How quantifying the shape of stories predicts their success. *Proc. Natl. Acad. Sci. United States Am.* **118** (2021).
30. YR Tausczik, JW Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
31. M Brysbaert, AB Warriner, V Kuperman, Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46** (2014).
32. Kourken Michaelian, Episodic and semantic memory and imagination: The need for definitions. *The Am. journal psychology* **131**, 99–103 (2018).
33. MA Greenberg, CB Wortman, AA 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 arXiv:2108.07258 (2021).

35. P Laban, L Dai, L Bandarkar, MA 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, Online), pp. 1058–1064 (2021).

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, Online), pp. 7282–7296 (2021).

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. United States Am.* **117**, 10165–10171 (2020).

38. JW Pennebaker, RJ Booth, RL Boyd, ME Francis, Linguistic inquiry and word count: LIWC 2015 (2015).

39. T van Laer, J Edson Escalas, S Ludwig, EA van den Hende, What happens in vegas stays on TripAdvisor? a theory and technique to understand narrativity in consumer reviews. *The J. consumer research* **46**, 267–285 (2019).

40. M Sap, E Horvitz, Y Choi, NA Smith, JW Pennebaker, Recollection versus imagination: Exploring human memory and cognition via neural language models in *ACL*. (2020).

41. J Hale, A probabilistic earley parser as a psycholinguistic model in *NAACL-HLT*. (Association for Computational Linguistics), pp. 1–8 (2001).

42. R Levy, Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).

43. A Goldstein, et al., Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines in *bioRxiv*. (Cold Spring Harbor Laboratory), (2021).

44. A Radford, et al., Language models are unsupervised multitask learners. unpublished (2019).

45. J Gordon, B Van Durme, Reporting bias and knowledge acquisition in *Proceedings of the 2013 workshop on Automated knowledge base construction*. pp. 25–30 (2013).

46. DL Schacter, DR Addis, The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philos. Transactions Royal Soc. B: Biol. Sci.* **362**, 773–786 (2007).

47. A Grysman, JA Hudson, The self in autobiographical memory: effects of self-salience on narrative content and structure. *Memory* **19**, 501–513 (2011).

48. CA Kurby, JM Zacks, Segmentation in the perception and memory of events. *Trends cognitive sciences* **12**, 72–79 (2008).

49. PW Thorndyke, Cognitive structures in comprehension and memory of narrative discourse. *Cogn. psychology* **9**, 77–110 (1977).

50. W Kintsch, E Greene, The role of culture-specific schemata in the comprehension and recall of stories. *Discourse processes* **1**, 1–13 (1978).

51. LR Squire, Two forms of human amnesia: An analysis of forgetting. *J. Neurosci.* **1**, 635–640 (1981).

52. A Piper, RJ So, D Bamman, Narrative theory for computational narrative understanding in *EMNLP*. (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic), pp. 298–311 (2021).

53. T Underwood, *The Invention of Historical Perspective.* (Stanford University Press), pp. 55–80 (2013).

54. T Underwood, Machine learning and human perspective. *PMLA/Publications Mod. Lang. Assoc. Am.* **135**, 92–109 (2020).

55. EJ Marsh, Retelling is not the same as recalling: Implications for memory. *Curr. directions psychological science* **16**, 16–20 (2007).

56. J Devlin, MW Chang, K Lee, K Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding in *NAACL*. (2019).

57. A Jafarpour, EA Buffalo, RT Knight, AG Collins, Event segmentation reveals working memory forgetting rate (2019).
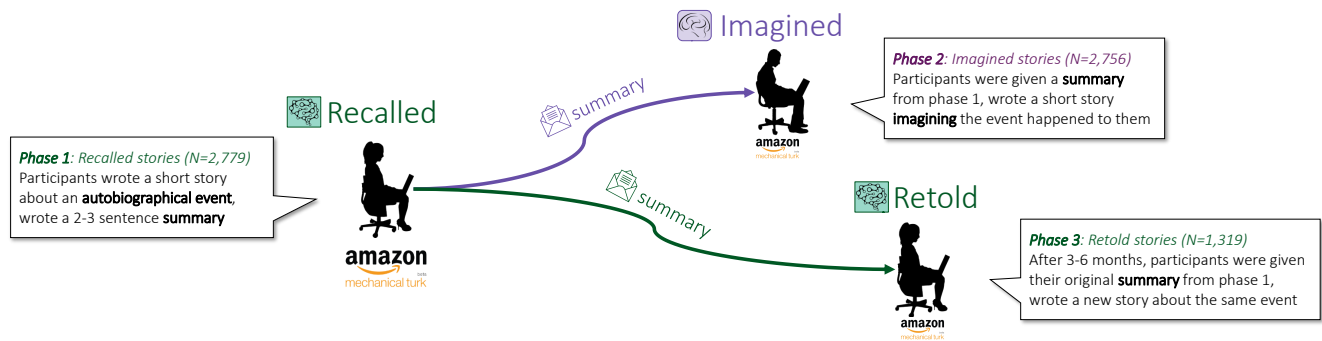
Fig. 5. Data collection pipeline for constructing HIPPOCORPUS.

A. Further Data Collection Details

A. HIPPOCORPUS. The stories in HIPPOCORPUS were collected from crowdworkers in three phases (see Figure 5). In phase one, a set of crowdworkers wrote stories about memorable events they had experienced in the recent past (3-6 months) and summarized their story in one to three sentences (N=2,779 *recalled stories*, written by 2662 authors[‡]). The summary of the story that was provided by the author was used as the *topic* of the story. An example of a topic is: "*My daughter and her husband announced the were expecting their second child. While on a camping trip she feared that she might be having a miscarriage only to learn that she was having twins.*"

In the second phase, we used the story topic after 3-6 month of writing the original memory, to ask a subset of the authors to retell the autobiographical story (N=1,319 *retold stories*). In the third phase, we provided the story topics to another set of crowdworkers and asked them to write *imagined stories* as if the event in the summary had happened to them (N=2,756 imagined stories, written by N=1434 authors[§]). During the recalled and retold storywriting tasks, we also asked workers the time elapsed since they experience 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 approved by the Microsoft Research Institutional Review Board (IRB) and sought explicit informed consent from participants. See http://aka.ms/hippocorpus for the data collection crowdsourcing templates.

B. Event annotations. To test the hypothesis that the sequentiality of narratives is associated with the number of events contained in stories, we ran an annotation task and analyzed the number of events in a random selection of $N = 240$ stories of the HIPPOCORPUS. The subset consisted of a triple of imagination, recalled, and retold stories for each of 80 topics. In this subset, the autobiographical stories on the same topic are written by the same person, to keep the author's schematic knowledge of the topic constant. Also, each triple of story types is annotated by the same participants ($n = 8$). Keeping the annotators within a topic constant allowed control of the variability in schematic information and the individual difference in event segmentation (18, 57).

In the annotation task, participants when through three stories, one sentence at a time, and indicated if the story sentences mark a start of a new event. We specifically asked the annotators to differentiate whether a new event is *minor* or *major* and, for 60 topics, we additionally asked if events are *expected* or *unexpected*. Given that the saliency of events can vary (18), participants were instructed to use their interpretation of what constitutes a major or minor event and if the events are expected or surprising. The order of the type of story posed for annotation was randomized.

This data collection effort was also approved by the Microsoft Research IRB.

C. Participants. We recruited a diverse group of story authors. The participants' age ranged between 18 and 55 years old (M = 33.6, SD = 10.5) and were 47% male, 52% female, <1% non-binary, and <1% other. They were 73.7% White, 10.1% Black, 5.2% Asian, 6.1 % Hispanic, 0.8% Native American, 0.7% Indian, 0.3% Middle Eastern, 0.2% Islander, 2.7% Other, and 0.7 % unidentified (this data has previously been published in 24). 189 participants annotated events in a selection of stories (18-55 years old (M=37, SD=10.6); 53% men, 46% women, and 1% unidentified; 75.7% white, 6.3% Asian, 5.8% Black, 4.2% Hispanic, 0.5% Indian, 0.5% Native American, 5.3% other, 1.6% unidentified). All studies were conducted on Amazon Mechanical Turk. The procedures were approved by Microsoft's ethical review board. All participants gave written informed consent before participation and were compensated.

B. Lexicon-based narrative measures: results

To complement the sequentiality and realis analyses, we used several lexicon-based measures to investigate the stylistic and content differences in imagined and autobiographical stories, as well as in sentences with minor, major, or no events. For each story or sentence, LIWC lexicon scores are computed as the proportion of words in that story or sentence that appear in LIWC categories (38). For the weighted concreteness lexicon (31), which contains a concreteness rating between 0–1 for 39,000 words, we compute a story's or a sentence's concreteness score as the average rating per word in the story or sentence.

We list the LIWC and concreteness scores, comparing proportions in imagined, recalled, and retold stories (Table 1) and the proportions in sentences with major, minor, or no events (Table 2).

As a way to verify the validity of these categories in our corpus (37), we list the top five most frequent words in our two corpora. Additionally, we also compute the coverage of each category, i.e., the proportion of words in our corpora that belong to each category.

[‡]2550 of the authors wrote one autobiographical story, 107 wrote two stories, and 5 wrote three stories

[§]As participation in this task was not restricted, 1,072 wrote one story, 311 between two and five stories, and 30 between six and ten stories, and 21 workers more than ten stories (following a Zipfian distribution between 11 and 92).

Variable	Recalled	retold	Imagined	Rrl	Rr	rl	RI	Coverage (%)	Top words
Concreteness	**0.321**	**0.322**	**0.317**	*		‡	†	87.1	the, to, and, a, was
Function words	**9.140**	**9.132**	**8.157**	*	+	‡	†	58.6	i, the, to, and, a
Total pronouns	2.860	2.841	2.660					18.6	i, my, it, that, we
Personal pronouns	2.049	2.020	1.914					13.4	i, my, we, me, he
1st pers singular	1.172	1.109	1.122					7.6	i, my, me, i'm, myself
1st pers plural	**0.310**	**0.334**	**0.277**	*		‡		2.0	we, our, us, we've, we're
2nd person	0.033	0.030	0.034					0.2	you, your, you're, yourself, you've
3rd pers singular	0.430	0.433	0.376					2.8	he, she, her, him, his
3rd pers plural	0.104	0.115	0.104					0.7	they, them, their*, themselves, they'd
Impersonal pronouns	0.810	0.820	0.745					5.2	it, that, this, what, thing*
Articles	**1.082**	**1.073**	**0.910**	*		‡	†	6.8	the, a, an
Prepositions	**2.197**	**2.200**	**1.898**	*	+	‡	†	13.9	to, of, in, for, with
Auxiliary verbs	1.389	1.396	1.294					9.0	was, had, have, is, be
Common adverbs	**0.884**	**0.905**	**0.837**	*	+			5.8	so, about, when, there, just
Conjunctions	**1.098**	**1.084**	**0.926**	*		‡	†	6.9	and, so, but, as, when
Negations	**0.218**	**0.219**	**0.216**	*			†	1.5	not, didn't, never, no, don't
Grammar Other									
Regular verbs	**2.833**	**2.841**	**2.645**	*	+			18.5	was, had, have, is, be
Adjectives	**0.775**	**0.772**	**0.702**	*		‡		5.0	as, after, more, few, new
Comparatives	0.376	0.377	0.333					2.4	as, like, after, more, before
Interrogatives	0.188	0.191	0.172					1.2	when, what, how, who, which
Numbers	**0.148**	**0.128**	**0.121**	*	+		†	0.9	one, first, two, three, once
Quantifiers	**0.387**	**0.409**	**0.347**	*	+	‡		2.5	all, some, more, much, few
Affect Words	0.796	0.792	0.776					5.2	good, well, great, happy, best
Positive emotion	0.527	0.537	0.529					3.5	good, well, great, happy, best
Negative emotion	0.258	0.244	0.236					1.7	bad, lost, pain, stress*, sad
Anxiety	0.063	0.061	0.055					0.4	stress*, nervous, worried, scared, struggl*
Anger	0.039	0.039	0.035					0.3	fight*, angry, attack*, frustrat*, kill*
Sadness	0.072	0.066	0.063					0.5	lost, sad, hurt*, alone, miss
Social Words	**1.732**	**1.753**	**1.564**	*		‡		11.2	we, he, she, her, they
Family	0.214	0.219	0.186					1.4	family, husband*, wife*, daughter*, son
Friends	0.074	0.068	0.071					0.5	friends, friend, guy*, girlfriend*, boyfriend*
Female referents	0.291	0.287	0.250					1.9	she, her, wife*, daughter*, sister*
Male referents	0.307	0.317	0.271					2.0	he, him, his, husband*, son
Cognitive Processes	**1.635**	**1.670**	**1.592**	*	+		†	10.9	but, all, not, would, some
Insight	**0.349**	**0.349**	**0.347**	*			†	2.3	know, felt, think, decided, feel
Cause	0.213	0.215	0.186					1.4	because, how, made, since, make
Discrepancies	**0.212**	**0.219**	**0.227**	*		‡	†	1.5	would, could, if, wanted, want
Tentativeness	**0.323**	**0.339**	**0.315**	*	+		†	2.2	some, or, if, lot, something*
Certainty	**0.240**	**0.238**	**0.241**	*			†	1.6	all, never, always, everyone*, everything*
Differentiation	**0.411**	**0.428**	**0.392**	*	+		†	2.7	but, not, really, or, didn't
Perceptual Processes	0.383	0.377	0.348					2.5	see, felt, said, feel, experienc*
Seeing	0.155	0.158	0.141					1.0	see, saw, looked, looking, beautiful
Hearing	0.083	0.080	0.077					0.5	said, say, phone*, laugh*, heard
Feeling	0.118	0.111	0.107					0.8	felt, feel, hard, feeling, pain
Biological Processes	0.353	0.343	0.314					2.3	life, love, food*, hospital*, doctor*
Body	0.083	0.073	0.068					0.5	eye*, heart, head, face, sleep
Health/illness	0.158	0.150	0.132					1.0	life, hospital*, doctor*, live, pain
Sexuality	0.008	0.006	0.007					0.0	pregnan*, passion*, sex, breast*, sexual*
Ingesting	0.094	0.101	0.093					0.6	food*, drink*, dinner*, water, restau*
Core Drives and Needs	**1.441**	**1.458**	**1.315**	*		‡		9.3	we, up, our, get, got
Affiliation	**0.645**	**0.659**	**0.582**	*				4.1	we, our, family, us, friends
Achievement	0.247	0.248	0.222					1.6	first, work, able, best, better
Power	0.324	0.315	0.287					2.1	up, over, down, best, help
Reward focus	0.275	0.279	0.274					1.8	get, got, good, took, great
Risk/prevention focus	0.064	0.067	0.063					0.4	bad, problem*, stop, wrong, difficult
Time orientation									
Past focus	**1.486**	**1.422**	**1.289**	*		‡	†	9.4	was, had, were, been, got
Present focus	**1.034**	**1.080**	**1.072**	*	+		†	7.0	have, is, be, get, do
Future focus	0.163	0.166	0.165					1.1	will, going, then, hope, expect*
Relativity	**2.657**	**2.601**	**2.271**	*		‡	†	16.8	in, on, at, time*, up
Motion	**0.456**	**0.456**	**0.383**	*		‡	†	2.9	went, go, going, came, car
Space	**1.096**	**1.110**	**0.944**	*	+	‡	†	7.0	in, on, at, up, out
Time	**1.171**	**1.103**	**0.998**	*		‡	†	7.3	time*, day*, when, after, month*
Personal Concerns									
Work	0.290	0.288	0.256					1.8	work, job*, school, doctor*, pay*
Leisure	0.229	0.231	0.207					1.5	family, fun, party*, game*, drink*
Home	0.134	0.130	0.114					0.8	family, home, house, room, door*
Money	0.104	0.110	0.093					0.7	money*, spent, pay*, spend, store
Religion	0.020	0.016	0.018					0.1	god, church*, bless*, spirit*, christmas*
Death	0.022	0.023	0.017					0.1	died, death*, funer*, kill*, dead
Informal Speech	0.051	0.050	0.050					0.3	well, anyway*, cool, okay, yes
Swear words	0.005	0.005	0.006					0.0	freak*, hell, crap, shit*, damn*
Netspeak	0.004	0.004	0.004					0.0	gonna, app, kinda, lol, o
Assent	0.014	0.013	0.015					0.1	cool, okay, yes, absolutely, ok
Nonfluencies	0.023	0.023	0.020					0.1	well, oh, er, sigh, ugh
Fillers	0.003	0.003	0.004					0.0	anyway*, blah, rr*, dunno, y'kn*

Table 1. Average lexicon scores for the three story types in HIPPOCORPUS (recalled (R), retold (r), and imagined (I)), along with significance values of the three-way and pairwise differences. The ∗,+,†,‡ symbols denote p-values <0.05 after Bonferroni correction. "Coverage" indicates the percentage of the total number of words in the HIPPOCORPUS that are in the lexicon category ("Variable"), with the five most common words listed under "Top words". We bold variables with significant three-way differences.

Variable	Major	minor	noEvent	Mmn	Mm	mn	Mn	Coverage (%)	Top words
Concreteness	**0.333**	**0.338**	**0.311**	*		‡	†	86.8	the, to, and, was, a
Function words	**8.725**	**7.652**	**8.144**	*		‡	†	58.7	i, the, to, and, was
Total pronouns	**2.776**	**2.421**	**2.647**	*			†	18.9	i, my, it, that, we
Personal pronouns	2.102	1.935	1.787					13.4	i, my, we, me, her
1st pers singular	1.263	1.084	1.042					7.9	i, my, me, i'm, myself
1st pers plural	**0.248**	**0.382**	**0.241**	*	+	‡		1.8	we, our, us, we've, we'd
2nd person	0.024	0.006	0.038					0.2	you, your, you're, yourself, you've
3rd pers singular	0.480	0.362	0.350					2.8	her, she, he, him, his
3rd pers plural	0.087	0.101	0.116				†	0.8	they, them, their*, themselves, they've
Impersonal pronouns	**0.673**	**0.486**	**0.860**	*		‡	†	5.5	it, that, this, thing*, what
Articles	**1.100**	**1.107**	**0.837**	*		‡	†	6.6	the, a, an
Prepositions	**2.231**	**2.056**	**1.785**	*		‡	†	13.7	to, of, in, for, with
Auxiliary verbs	**1.216**	**0.874**	**1.401**	*	+	‡	†	9.3	was, had, have, is, were
Common adverbs	**0.779**	**0.615**	**0.887**	*		‡	†	5.9	so, about, when, just, there
Conjunctions	**0.952**	**0.958**	**0.974**	*			†	6.9	and, so, but, as, when
Negations	**0.178**	**0.051**	**0.244**	*	+	‡	†	1.5	not, didn't, never, no, wasn't
Grammar Other									
Regular verbs	**2.673**	**2.272**	**2.645**	*		‡	†	18.6	was, had, have, is, were
Adjectives	0.722	0.587	0.715					5.0	as, after, more, great, happy
Comparatives	0.361	0.239	0.349					2.4	as, like, after, more, before
Interrogatives	0.197	0.149	0.168					1.2	when, what, how, who, where
Numbers	**0.198**	**0.087**	**0.098**	*	+		†	0.9	one, first, two, three, once
Quantifiers	**0.298**	**0.343**	**0.342**	*			†	2.4	all, some, more, much, few
Affect Words	**0.647**	**0.559**	**0.824**	*		‡	†	5.3	good, great, well, happy, love
Positive emotion	**0.408**	**0.458**	**0.601**	*			†	3.8	good, great, happy, well, love
Negative emotion	0.227	0.101	0.213			‡		1.5	bad, pain, sad, problem*, miss
Anxiety	0.032	0.031	0.057				†	0.3	upset, worried, nervous, stress*, struggl*
Anger	0.046	0.011	0.028					0.2	frustrat*, kill*, cheat*, fight*, argu*
Sadness	0.058	0.042	0.064					0.4	sad, miss, tears, alone, lost
Social Words	**1.816**	**1.702**	**1.462**	*				11.2	we, her, she, he, they
Family	**0.330**	**0.230**	**0.178**	*			†	1.6	family, son, parent*, daughter*, husband*
Friends	0.090	0.087	0.054				†	0.5	friend, friends, boyfriend*, guy*, girlfriend*
Female referents	**0.398**	**0.256**	**0.228**	*			†	2.0	her, she, daughter*, wife*, sister*
Male referents	0.326	0.261	0.235					1.9	he, him, his, son, husband*
Cognitive Processes	**1.319**	**0.910**	**1.665**	*		‡	†	10.7	but, all, not, would, some
Insight	**0.293**	**0.197**	**0.339**	*		‡	†	2.2	know, think, felt, feel, decided
Cause	**0.165**	**0.129**	**0.198**	*			†	1.3	how, because, made, make, since
Discrepancies	**0.163**	**0.126**	**0.239**	*		‡	†	1.5	would, could, wanted, if, want
Tentativeness	**0.235**	**0.222**	**0.321**	*			†	2.1	some, or, if, lot, something*
Certainty	**0.162**	**0.132**	**0.254**	*		‡	†	1.5	all, never, everyone*, always, everything*
Differentiation	**0.359**	**0.146**	**0.430**	*	+	‡	†	2.7	but, not, really, didn't, other
Perceptual Processes	0.386	0.382	0.357					2.6	see, felt, said, feel, eye*
Seeing	0.180	0.216	0.170					1.3	see, eye*, looked, saw, seen
Hearing	0.102	0.101	0.061					0.5	said, say, phone*, laugh*, hear
Feeling	0.089	0.059	0.101					0.7	felt, feel, hard, pain, feeling
Biological Processes	**0.416**	**0.385**	**0.285**	*			†	2.3	life, eye*, love, food*, doctor*
Body	**0.112**	**0.053**	**0.057**	*			†	0.5	eye*, head, knee*, face, heart
Health/illness	0.180	0.093	0.125					1.0	life, doctor*, live, surger*, pain
Sexuality	**0.019**	**0.000**	**0.006**	*			†	0.1	pregnan*, prostat*, passion*, virgin*, screw*
Ingesting	**0.112**	**0.222**	**0.078**	*	+	‡		0.7	food*, dinner*, cake*, water, restau*
Core Drives and Needs	**1.454**	**1.475**	**1.225**	*		‡		9.3	we, up, our, family, got
Affiliation	**0.666**	**0.792**	**0.545**	*		‡		4.3	we, our, family, us, together
Achievement	0.268	0.174	0.209					1.6	first, able, work, best, proud
Power	0.334	0.323	0.275					2.1	up, over, down, help, best
Reward focus	0.257	0.244	0.255					1.8	got, get, good, great, took
Risk/prevention focus	0.055	0.022	0.052					0.4	bad, problem*, stop, wrong, stopped
Time orientation									
Past focus	**1.683**	**1.354**	**1.210**	*			†	9.6	was, had, were, been, got
Present focus	**0.784**	**0.598**	**1.155**	*		‡	†	7.1	have, is, be, get, go
Future focus	**0.132**	**0.104**	**0.174**	*			†	1.1	going, will, then, expect*, hope
Relativity	**2.909**	**2.306**	**2.069**	*			†	16.5	in, on, time*, at, out
Motion	**0.504**	**0.500**	**0.357**	*		‡	†	2.9	went, go, going, came, trip
Space	**1.177**	**1.051**	**0.865**	*		‡	†	6.9	in, on, at, out, up
Time	**1.292**	**0.812**	**0.905**	*	+		†	7.1	time*, when, day*, after, month*
Personal Concerns									
Work	0.319	0.219	0.216					1.7	work, job*, graduat*, school, doctor*
Leisure	**0.237**	**0.348**	**0.180**	*	+	‡		1.5	family, fun, party*, celebrat*, garden*
Home	**0.145**	**0.180**	**0.097**	*		‡		0.8	family, home, house, garden*, room
Money	**0.104**	**0.160**	**0.069**	*		‡		0.6	spent, income*, money*, bought, spend
Religion	0.007	0.008	0.013					0.1	god, zion*, church*, ritual*, prayed
Death	0.031	0.014	0.012				†	0.1	death*, died, funer*, kill*, die
Informal Speech	0.045	0.039	0.053					0.4	well, okay, yes, anyway*, cool
Swear words	0.010	0.000	0.004					0.0	crap, freak*, hell, sucks, screw*
Netspeak	0.003	0.003	0.009					0.0	kinda, o, lil, ha, haha*
Assent	0.011	0.008	0.014					0.1	okay, yes, cool, awesome, absolutely
Nonfluencies	0.020	0.025	0.025					0.2	well, er, oh, ugh, sighed
Fillers	0.001	0.003	0.003					0.0	anyway*

Table 2. Average lexicon scores for the three event types (major (M), minor (m), and no event (n)) in the annotated subset, along with significance values of the three-way and pairwise differences. The ∗,+,†,‡ symbols denote p-values <0.05 after Bonferroni correction. "Coverage" indicates the percentage of the total number of words in the event-annotated sentences that are in the lexicon category ("Variable"), with the five most common words listed under "Top words". We bold variables with significant three-way differences.