Computational Lens on Cognition: Study Of Autobiographical Versus Imagined Stories With Large-Scale Language Models

Maarten Sap[*]
msap@cs.washington.edu
University of Washington
Allen Institute for AI
Seattle, WA, USA

Anna Jafarpour[*]
annaja@uw.edu
University of Washington
Seattle, WA, USA

Yejin Choi
yejin@cs.washington.edu
University of Washington
Allen Institute for AI
Seattle, WA, USA

Noah A. Smith
nasmith@cs.washington.edu
University of Washington
Allen Institute for AI
Seattle, WA, USA

James W. Pennebaker
pennebaker@utexas.edu
University of Texas at Austin
Austin, TX, USA

Eric Horvitz
horvitz@microsoft.com
Microsoft
Redmond, WA, USA

Abstract

Lifelong experiences and learned knowledge lead to shared expectations about how common situations tend to unfold. Such knowledge enables people to interpret story narratives and identify salient events effortlessly. We study differences in the narrative flow of events in autobiographical versus imagined stories using GPT-3, one of the largest neural language models created to date. The diary-like stories were written by crowdworkers about either a recently experienced event or an imagined event on the same topic. To analyze the narrative flow of events of these stories, we measured sentence *sequentiality*, which compares the probability of a sentence with and without its preceding story context. We found that imagined stories have higher sequentiality than autobiographical stories, and that the sequentiality of autobiographical stories is higher when they are retold than when freshly recalled. Through an annotation of events in story sentences, we found that the story types contain similar proportions of major salient events, but that the autobiographical stories are denser in factual minor events. Furthermore, in comparison to imagined stories, autobiographical stories contain more concrete words and words related to the first person, cognitive processes, time, space, numbers, social words, and core drives and needs. Our findings highlight the opportunity to investigate memory and cognition with large-scale statistical language models.

Keywords: Natural language processing | autobiographical memory | memory consolidation | imagination | deep neural network | pretrained language models

Introduction

When we tell a story, we weave together sets of events to form a coherent narrative (1, 2, 24). The narrative flow of events is informed by our recollection of experiences from episodic memory (10, 11, 44) as well as our common knowledge about prototypical sequences of events, also referred to as *schema* (1, 3, 16, 19, 22, 37). For example, telling an *imagined* story about a friend's wedding relies on common knowledge about the schema of how a wedding unfolds. In contrast, a recalled story about an *autobiographical* memory of a friend's wedding involves recollection of episodic details about events experienced in addition to knowledge of wedding schema (13). Studies have found that, as time passes since events have been experienced, memories are consolidated and schematized into more abstract, semantic, and "gist-like" versions (9, 39, 47). Such changes may shape retold autobiographical stories to be more similar to imagined stories.

A cornerstone of storytelling is the incorporation of salient *events.* These events are often marked by surprising or expected shifts in character, cause, goal, location, or circumstances within the narrative flow of a story (50). People can identify salient events in stories with ease based on familiar understandings of narrative flow (51). Such events often stand out as particularly memorable (12, 34). Additionally, the saliency of events can vary in magnitude (21), as people can consider salient events as ranging from major plot twists to minor detailed events.

We studied the flow of sentences in narratives with statistical models, focusing on identifying the differences between stories that are based on autobiographical memory versus processes of imagination. In pursuit of this goal, we previously created Hippocorpus, a dataset of 7,000 diary-like short stories about memorable life experiences collected through crowdsourcing (36). By design, 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.

[*] Equal contribution

In addition to the characterization of the flow of a narrative, we investigated the nature and density of salient events in stories. In support of this analysis, we collected sentence-level event annotations for a subset of 240 stories from Hippocorpus (36). We asked crowdworkers to identify the presence of salient events and to annotate whether the events are major versus minor and surprising versus expected. We further employed an automated proxy for identifying events via use of a predictive model trained on an annotated corpus tagged for *realis* terms (38). Realis terms are non-hypothetical references to concrete events that took place (e.g., "she tripped," as opposed to "she feared tripping"). We also noted average numbers of words in stories falling in psychologically related categories using the Linguistic Inquiry Word Count (LIWC) (41), and measured the average concreteness level of words using the lexicon (6).

To study the narrative flow of events in the three story types, we employed *sequentiality*, a statistical measure of the likelihood of sentences with respect to their preceding context (36). Sequentiality is distinct from earlier characterizations used to explore narrative flow, which either focused on detecting event words from sentences (29, 38) or tracking attributes over time in stories (e.g., sentence sentiment; dictionaries; sentence embeddings; 4, 33, 43). Instead, our method leverages transformer-based (48) large-scale neural language models, trained to predict the probabilities of words in their preceding context. We consider events at the granularity of sentences as in prior work (25, 49), and average sequentiality over a series of sentences to examine stories.

We drew inferences from GPT-3, one of the largest neural language models created to date, scaling up from our previous investigations (36). GPT-3 was trained on hundreds of billions of words from 570 gigabytes of general English internet documents (including news articles, fiction books, Wikipedia, etc.; 5). Harnessing inferences from GPT-3, sequentiality was computed as the difference between the log-likelihoods of consecutive sentences conditioned on the story topic and prior sentences and the log-likelihoods of sentences conditioned only on the topic. As one of the largest and most sophisticated language models in existence, GPT-3's ability to model language and predict ordering of sentences (5) suggests we can consider its inferences about the likelihood of sequences of words as a proxy for the degree to which a story follows commonsense or general expectations of how narratives unfold.

We hypothesized that autobiographical and imagined stories differ in their sequentiality and their event proportions, based on the intuition that imagined stories are more likely to be constructed sequentially with a focus on chaining events, whereas autobiographical stories are based on the specific details of experienced events that have been encoded in episodic memory (11, 17). We also hypothesized that there would be differences in sequentiality and event density for stories that are retold after a period of time versus freshly recalled memories, due to memories being consolidated and narrativized as time goes on (39). Finally, we predicted that the proportion of salient events differs in autobiographical and imagined stories, and that we would find differences in the types of words used to refer to salient events.

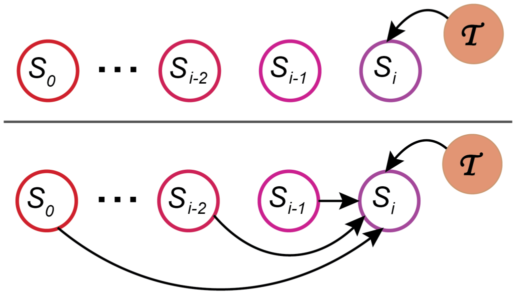

Figure 1. Graphical models depicting the two components of sequentiality. Sequentiality reflects the probabilistic relationship among consecutive sentences in a story, and is computed as the difference between the log-likelihood of a sentence conditioned only on the story topic (i.e., topic-centric; top row) and the log-likelihood of that sentence conditioned on both the story topic and all preceding sentences (bottom row depicts the full story history). The log-likelihood of a sentence given a topic or topic and prior sentences is provided by the GPT-3 neural language model.

In summary, we harnessed a large-scale neural language model to probe hypotheses about the cognitive processes of constructing stories. We considered differences in the narrative flow, event density, and word usage across (1) stories based on freshly recalled autobiographical experiences, (2) retellings of the stories following a delay of three to six months, and (3) imagined stories matched to the topics of the autobiographical stories. To examine the narrative event flow of stories, we employed sequentiality (36), which measured the likelihood of story sentences with respect to their context using the GPT-3 large-scale language model trained on (5). We applied our methods to study the differences in autobiographical versus imagined stories from the Hippocorpus (36), as well as uncover shifts in attributes of autobiographical stories when they are retold months after they were first written down, shortly after being experienced. Such shifts could provide traces of processes of memory consolidation and increasing reliance on commonsense schemas with the passage of time since events were experienced.

Sequentiality with Large-scale Language Models

We harnessed *sequentiality* to quantify how much a story and its sentences deviate from the expected narrative flow given the story's topic using a large-scale neural language model

(LM). Formally, a language model estimates the probability of a word w_i given its preceding context $w_{0:i-1}$: $p_{\text{LM}}(w_i \mid w_{0:i-1})$. Large-scale neural language models are large multilayer neural networks that are trained on the language modeling task in a self-supervised fashion, i.e, its parameters are estimated using large amounts of unlabeled text data. These large-scale language models can be thought of as encoding the expected narrative flow of events (5, 27, 32), based on the assumption that the language for describing a common scenario and the stereotypical flow of events are similar (35).

Using large-scale neural language models, we computed sequentiality as follows. Given a story written about a topic $\mathcal{T}$ (represented as a two-three sentence summary), the sequentiality of a story sentence is computed as the difference in log-likelihood of that sentence under two generative models, portrayed in Fig. 1. The *topic-driven* generative model assumes that every sentence follows after the main topic of the story ($\mathcal{T}$). Conversely, the *contextual* model views a story as a main topic followed by a linear chain of consecutive sentences.[1] We then define a story's sequentiality as the average sequentiality of its sentences.

Formally, a sentence's sequentiality $c(s_i, h)$ is computed as the difference in negative log-likelihood (NLL) of a sentence s_i conditioned on words from the h preceding sentences $(s_{i-h:i-1})$ and words from the story topic $\mathcal{T}$, versus just the story topic (normalized by sentence length $|s_i|$ to account for length differences; 36):[2]

$$c(s_i, h) = -\frac{1}{|s_i|}\left[\log p_{\text{LM}}(s_i \mid \mathcal{T}) - \log p_{\text{LM}}(s_i \mid \mathcal{T}, s_{i-h:i-1})\right]$$

Here, the log likelihood of a sentence s_i in a context is measured as the sum of log probabilities of each of the words w_t in s_i: $\log p(s_i \mid C) = \sum_t \log p_{\text{LM}}(w_t \mid C, w_0, ..., w_{t-1})$ where C is the conditioning context of the sentence (words in the preceding sentences and/or story topic).

Higher values of sequentiality for sentences suggests that the sentences follow the expectations given the context of the story and topic, whereas lower values suggest that sentences deviate more from expectation. We note that taking the ratio of likelihoods provided by the contextual and topic-driven generative models allows for comparisons of sequentiality across topics, as both components of sequentiality are conditioned on the same topic.

In analyses, we studied the sensitivity of the computed sequentiality to the size of the preceding context (the history h), from a single preceding sentence ($h = 1$) to considering all sentences in the preceding context (h = full). The likelihoods of words and sentences were inferred with GPT-3. We used the story summaries provided by the storyteller as the topics $\mathcal{T}$. We also considered the topic-driven likelihood of sentences, computed by conditioning the sentences of stories only on the topic ($h = 0$); we report the negative log-likelihood of sentences, $\text{NLL}_{\mathcal{T}} = -\frac{1}{|s_i|} \log p_{\text{LM}}(s_i \mid \mathcal{T})$.

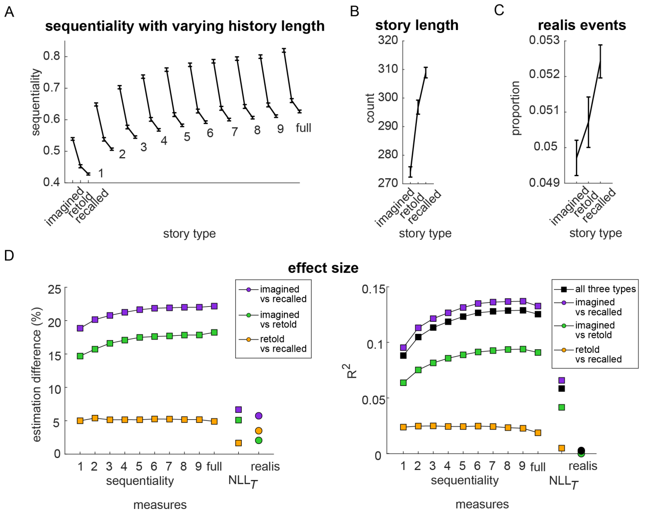

Figure 2. Differences in sequentiality in recalled, retold, and imagined stories. (A) Mean sequentiality values with varying history lengths (h = 1 to h = full story length) are different across the story types. Imagined stories have higher average sequentiality than autobiographical stories, and retold 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 likelihoods of sentences, $\text{NLL}_{\mathcal{T}}$, and the proportion of realis across story types.

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 reported 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.

[1] Note that the contextual model is a sentence-level version of the word-level *surprisal* as defined by expectation theory (18, 28).

[2] This measure is equivalent to the average pointwise mutual information (PMI; 7) between each word in the sentence and the preceding sentences, conditioned on the words in the topic.

Imagined stories flow in a more expected manner than autobiographical stories. The comparison 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. We observe a pattern of higher NLL$_{\mathcal{T}}$ (indicating lower topic-driven likelihood) for sentences of imagined stories versus recalled stories when conditioning only on the topic sentences, suggesting that the sentences of imagined stories on average have weaker links to the topic than sentences of autobiographical stories. However, the differences across story types measured with sequentiality (with increasing history size) dominate differences detected with the overall increased topic-centricity of autobiographical stories with the differences in story length and the number of realis events have lower effect sizes and R^2 (Fig. 2D).

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 suggests that autobiographical memories in retold stories are more consolidated and narrativized. Although the assessed subjective frequency of recalling or retelling autobiographical memories is not associated with sequentiality, sequentiality negatively correlates 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 lexica (6, 30), is different across story types ($p < 0.001$; supplementary Tab. 1a) 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 ($p > 0.1$). 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. 1a).

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 lexica at the sentence level.

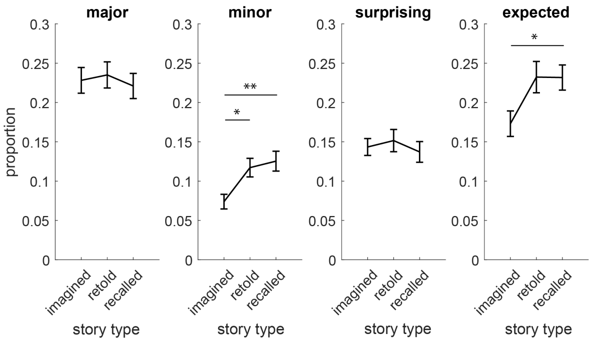

Figure 3. Event annotation across the stories The mean and S.E.M. of the proportion of annotations (from left: major, minor, surprising, and expected events) in the imagined stories and retold and recalled autobiographical stories. (* $P < 0.05$, ** $P < 0.01$)

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 ($p = 0.025$) 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 events and concrete, emotional, and time-and-space words. We found a higher proportion of realis events in sentences with minor events than those with a major ($p < 0.001$) or no events ($p < 0.001$; Fig. 4C). Furthermore, we found numerous differences in proportions of cognition-related words across sentences with major, minor, or no event (see supplementary Table 1b). Using LWIC categories, we detected differences in word usage between sentences with salient events and no events. We detected limited differences between sentences annotated as containing minor and major 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.

Discussion

Large-scale deep neural network models have an extraordinary capacity to generate linguistic continuations of natural language prompts (5, 8). The models provide the probability of words given a context captured by preceded sentences that is similar to human predictions (14). Using the GPT-3 language model, we used *sequentiality* to quantify how much a story resonates with the expected or commonsense narrative for a story topic (Fig. 1). With the measure, we observed that a difference between experienced and imagined stories can be captured by differences in episodic details and differing reliance on schematic knowledge about how events in the stories should unfold (Fig. 2). Based on sequentiality differences, imagined stories have greater alignment with expectations and commonsense on the flow of sentences than autobiographical stories. Autobiographical stories contain more minor events than imagined stories (Fig. 3), and they tend to have higher proportions of first person references, contain more adjectives, conjunctions, quantifiers, and words referring to cognitive processes. Autobiographical stories also contain more concrete words and words referring to orientation in time with a focus on the past and present, as well as words relating to time, space, motion, and core drives and needs (supplementary Table 1a).

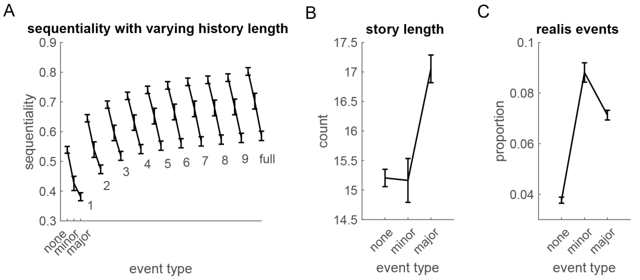

Figure 4. Sequentiality of sentences relative to event annotations (A) The average sequentiality with varying history is shown grouped by the event type. The sentences with no event (none) follow the narrative flow of the story topics more than sentences with major (all sequentiality history length) or minor events do (with sequentiality history of 1 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.

In all stories, storytellers appear to combine commonsense knowledge with references to major events. We found that major events are surprising sentences that tend to deviate from expectation, per likelihoods provided by the neural language model. They 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 1b). 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 are 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 are detectable 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. A subset of these observations has been previously reported in studies on detection of salient events (a.k.a., event boundaries; 25). 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 commonsense are generally shorter (23, 42).

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 affect the sequentiality. 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 commonsense and semantic knowledge of schema (1, 3, 39), especially since certain events may be forgotten (40).

Context, as formulated in the sequentiality, consists of the preceding sentences and the story topic. We observed that the extent to which sentences follow the direct preceding context versus story topic (measured with sequentiality) can be informative for detecting salient events in narratives (Fig. 4). Recently, a similar Bayesian model was proposed to detect the contextual transition to detect salient events in naturalistic videos (12). An advantage of a Bayesian model is its flexibility for accommodating the variety of the schema that can be attributed to a word. Bayesian models are versatile and have been developed to learn relational structures (26). For example, "broken glass" and "broken promise" are both likely, although with different probabilities given contextual information, whereas "broken moon" is unlikely in most contexts.

The likelihood of events in a context can be different across the topics. For example, driving on a highway for 30 miles has fewer expected events than a birthday party that has opportunities for details on whose birthday it was, where it took place, how the cake tasted, etc. We controlled for the variability across topics by quantifying sequentiality conditioned on story topics in both the topic-driven and contextual generative models. We observe that the magnitude of the difference between story types is greater using sequentiality than the NLL$_{\mathcal{T}}$ under just the topic-driven generative model. As such, sequentiality helps to minimize the variability in the twist and turns of the narrative flows that can be specific to the story topics, making the measure reliable in spite of the variation.

Besides the effect of variability in story topic that can affect the writing and reading of narratives, a prototypical schema of narrative flow can be different across social identities and cultures. For example, the schema for the progression of events on the evening of the Lunar New Year differs for people who celebrate the holiday and those who do not observe it. To control for such variability in understanding the narratives, we designed our salient event annotation task such that a full triple of imagined-recalled-retold stories about the same topic were annotated by the same person. Therefore, the experimental design controls for the diversity in knowledge across individual event annotators.

The sequentiality measure builds on the assumption that the language for describing common scenarios are similar (35) and that large-scale neural language models are learning the commonsense narrative flow of events (31). However, the extent to which the the language models learn the common flows of events is influenced by the knowledge contained their training data, which can disproportionately highlight noteworthy events rather than common events due to reporting bias (15). 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. We found that sequentiality shows similar differences across the three story types using language models trained on other data sets (36), such as OpenAI-GPT (trained on 5GBs of English fiction; 31) and GPT-2 (trained on 40GBs of news-like English text; 32).

In summary, we harnessed large-scale neural language models as novel tools for studying processes of recollection and imagination, as revealed in the narrative flow and event structure of autobiographical versus imagined stories. We used the language models to provide syntheses of community expectations about the likelihood of sequences of events. We aimed the inferences at better understanding processes of recall and imagination, including analyses of potential links between the recollection and imagination via studies of changes in recollection over time. We described how we harnessed the inferences about the likelihood of sentences in the context of a topic and prior sentences to compute the sequentiality of stories. We used this statistical measure of narrative flow to characterize differences in freshly recalled autobiographical stories, stories about the same experiences retold at a later time, and imagined stories on the same topic. We found that imagined stories have higher sequentiality than autobiographical stories, which we interpret as imagined stories being more aligned with the commonsense schemas of the narrative flow of events. We found that the narrative flow of retold stories has greater sequentiality than the fresher recollections, providing evidence of consolidation and narrativization of stories over time. In an analysis of events described in stories, we found that autobiographical and imagined stories contain a comparable proportion of events labeled as major and surprising by annotators. However, autobiographical stories contain higher numbers of factual events labeled as minor and expected. We found that autobiographical stories that are recalled fresh or retold after three to six months have comparable proportions of events.

We believe the methods and results provide new insights about memory and reasoning and that they light a path forward on opportunities for harnessing inferences from large-scale neural models to study human cognition, experience, and behavior. We see opportunities ahead to pursue answers to standing questions about memory and reasoning for individuals, groups, and for culture more broadly, including questions about the influences and interpretations of world events over time on fiction and non-fiction narrativizations (45, 46). Research directions forward include applying the results, methods, and measures in studies of recall, storytelling, lie-detection, false confessions, recovered memories, and the propagation and effects of misinformation.

Materials and Methods

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 36). 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.

Procedure

We previously created a large corpus of autobiographical, imagined, and retold stories, named Hippocorpus (36) to explore the use of large-scale neural language models as a tool for probing the differing cognitive processes employed in constructing narratives based on the recall of experienced stories versus the generation of imagined stories.

The stories in Hippocorpus were collected from crowdworkers in three stages. In the first stage, 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[3]). In the second stage, we provided the story summaries 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[4]). 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").

To address the hypothesis that the linearity of narratives reflects the number of events contained in stories, we analyzed the number of events in a random selection of N = 240 stories of the Hippocorpus. The stories consist of a triple of imagination, recalled, and retold stories for each of 80 topics. 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 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 (20, 21).

In the annotation task, participants read each story one sentence at a time, and they indicated if the 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 (21), 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.

The data as well as the story collection and event boundary annotation tasks are available at http://aka.ms/hippocorpus.

Data Analysis

Each collected story was segmented into sentences using a version of the NLTK sentence tokenizer adapted to avoid

[3] 2550 of the authors wrote one autobiographical story, 107 wrote two stories, and 5 wrote three stories

[4] 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).

splitting sentences into one-word sentences. The sequentiality was averaged across sentences in a story for story type difference analysis. We determined the difference in the mean of the values of interest that is assigned to the three story type using a factorial linear regression. The three story types are imagined stories, recalled, or retold memories. Besides the negative log likelihood (NLL), sequentiality with history length of 1 sentence to full story, realis, and the total number of words (story length), we evaluated the difference in the proportion of major or minor events across the story types. 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.

The main effect of grouping into the three story categories of imagination, recalled, and retold stories is evaluated before the pairwise comparison of the story types (imagination vs recall, imagination vs retold, and retold vs recall stories).

We tested if the salient sentences are less schematic than sentences with no events by comparing sentences according to the majoritarian annotation. This analysis is at the sentence level with 9412 major, 6835 minor, and 17477 no event annotation. We also tested the main effect of event annotation on the NLL, the sequentiality with history of 1 to the full story, as well as total words and realis. We used Bonferroni correction to adjust the significance threshold for multiple comparisons. All reported p-values are Bonferroni corrected. We employed a pairwise t-test for the posthoc tests.

Acknowledgements

Microsoft provided an internship for M.A.S, support for crowdworkers, and computing resources for running the GTP-3 neural language model on Hippocorpus. A.J. was supported by NIH Brain Initiative grant K99MH120048. We thank Paul Koch for his assistance with GPT-3 compute. We also thank Zhilin Wang and members of the Buffalo Lab at the University of Washington for valuable discussions.

References

[1] Frederic Charles Bartlett. 1932. *Remembering: A study in experimental and social psychology.* Cambridge University Press.

[2] John B Black and Hyman Bern. 1981. Causal coherence and memory for events in narratives. *Journal of Verbal Learning and Verbal Behavior* 20, 3 (June 1981), 267–275.

[3] Gordon H Bower, John B Black, and Terrence J Turner. 1979. Scripts in memory for text. *Cognitive psychology* 11, 2 (1979), 177–220.

[4] Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science advances* 6, 32 (Aug. 2020), eaba2196.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. (2020). unpublished.

[6] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46, 3 (2014).

[7] Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16, 1 (1990), 22–29.

[8] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. 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, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[9] David Clewett, Sarah DuBrow, and Lila Davachi. 2019. Transcending time in the brain: How event memories are constructed from experience. *Hippocampus* 29, 3 (2019), 162–183.

[10] Martin A. Conway, Alan F. Collins, Susan E. Gathercole, and Stephen J. Anderson. 1996. Recollections of true and false autobiographical memories. *Journal of Experimental Psychology: General* 125, 1 (1996).

[11] Martin A. Conway, Christopher W. Pleydell-Pearce, Sharron E. Whitecross, and Helen Sharpe. 2003. Neurophysiological correlates of memory for experienced and imagined events. *Neuropsychologia* 41, 3 (2003), 334–340.

[12] Nicholas T Franklin, Kenneth A Norman, Charan Ranganath, Jeffrey M Zacks, and Samuel J Gershman. 2020. Structured Event Memory: A neuro-symbolic model of event cognition. *Psychological Review* 127, 3 (2020), 327.

[13] Asaf Gilboa, R Shayna Rosenbaum, and Avi Mendelsohn. 2018. Autobiographical memory: From experiences to brain representations. *Neuropsychologia* 110 (Feb. 2018), 1–6.

[14] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2021. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. In *bioRxiv.* Cold Spring Harbor Laboratory. https://www.biorxiv.org/content/early/2021/03/19/2020.12.02.403477

[15] Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction.* 25–30.

[16] Arthur C Graesser, Scott P Robertson, and Patricia A Anderson. 1981. Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology* 13, 1 (Jan. 1981), 1–26.

[17] Melanie A. Greenberg, Camille B. Wortman, and Arthur A. Stone. 1996. Emotional expression and physical health: revising traumatic memories or fostering self-regulation? *Journal of Personality and Social Psychology* 71, 3 (Sept. 1996), 588–602.

[18] John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *NAACL-HLT.* Association for Computational Linguistics, 1–8.

[19] Ira E Hyman Jr and Elizabeth F Loftus. 1998. Errors in autobiographical memory. *Clinical psychology review* 18, 8 (1998), 933–947.

[20] Anna Jafarpour, Elizabeth A Buffalo, Robert T Knight, and Anne GE Collins. 2019. Event segmentation reveals working memory forgetting rate.

[21] Anna Jafarpour, Sandon Griffin, Jack J Lin, and Robert T Knight. 2019. Medial orbitofrontal cortex, dorsolateral prefrontal cortex, and hippocampus differentially represent the event saliency. *Journal of cognitive neuroscience* 31, 6 (2019), 874–884.

[22] Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95, 2 (1988), 163.

[23] Walter Kintsch and Edith Greene. 1978. The role of culture-specific schemata in the comprehension and recall of stories. *Discourse processes* 1, 1 (1978), 1–13.

[24] Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences* 12, 2 (Feb. 2008), 72–79.

[25] Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences* 12, 2 (2008), 72–79.

[26] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. 2018. Bayesian models of cognition. https://doi.org/10.1184/R1/6613682.v1

[27] Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. 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, 1058–1064. https://doi.org/10.18653/v1/2021.acl-short.134

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

[29] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[30] James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. Linguistic Inquiry and Word Count: LIWC 2015.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018). unpublished.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019). unpublished.

[33] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, 1 (2016), 31.

[34] Richárd Reichardt, Bertalan Polner, and Péter Simor. 2020. Novelty manipulations, memory performance, and predictive coding: The role of unexpectedness. *Frontiers in human neuroscience* 14 (2020), 152.

[35] David E Rumelhart and Andrew Ortony. 1977. The representation of knowledge in memory. *Schooling and the acquisition of knowledge* 99 (1977), 135.

[36] Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James W. Pennebaker. 2020. 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.

[37] Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum.

[38] Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. In *ACL*. https://www.aclweb.org/anthology/P19-1353

[39] Andrea Smorti and Chiara Fioretti. 2016. Why narrating changes memory: a contribution to an integrative model of memory and narrative processes. *Integrative Psychological and Behavioral Science* 50, 2 (2016), 296–319.

[40] LARRY R Squire. 1981. Two forms of human amnesia: An analysis of forgetting. *Journal of Neuroscience* 1, 6 (1981), 635–640.

[41] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.

[42] Perry W Thorndyke. 1977. Cognitive structures in comprehension and memory of narrative discourse. *Cognitive psychology* 9, 1 (1977), 77–110.

[43] Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. 2021. How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences of the United States of America* 118, 26 (June 2021).

[44] Endel Tulving. 1972. Episodic and semantic memory. *Organization of Memory* 1 (1972), 381–403.

[45] Ted Underwood. 2013. *The Invention of Historical Perspective*. Stanford University Press, 55–80.

[46] Ted Underwood. 2020. Machine Learning and Human Perspective. *PMLA/Publications of the Modern Language Association of America* 135, 1 (2020), 92–109.

[47] Marlieke T. R. van Kesteren, Dirk J. Ruiter, Guillén Fernández, and Richard N. Henson. 2012. How schema and novelty augment memory formation. *Trends in Neurosciences* 35, 4 (April 2012).

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.

[49] Jeffrey M. Zacks. 2020. Event Perception and Memory. *Annual Review of Psychology* 71, 1 (2020), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101 arXiv:https://doi.org/10.1146/annurev-psych-010419-051101 PMID: 31905113.

[50] Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. 2007. Event perception: a mind-brain perspective. *Psychological bulletin* 133, 2 (2007), 273.

[51] Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin* 123, 2 (1998), 162.

Table 1. Supplementary tables

Variable	*Recalled*	*retold*	*Imagined*	*p-values*			
				Rrl	*Rr*	*rI*	*RI*
Concreteness	**0.321**	**0.322**	**0.317**	*		‡	†
Function words	**9.140**	**9.132**	**8.157**	*	+	‡	†
Total pronouns	2.860	2.841	2.660				
Personal pronouns	2.049	2.020	1.914				
1st pers singular	1.172	1.109	1.122				
1st pers plural	**0.310**	**0.334**	**0.277**	*		‡	
2nd person	0.033	0.030	0.034				
3rd pers singular	0.430	0.433	0.376				
3rd pers plural	0.104	0.115	0.104				
Impersonal pronouns	0.810	0.820	0.745				
Articles	**1.082**	**1.073**	**0.910**	*		‡	†
Prepositions	**2.197**	**2.200**	**1.898**	*	+	‡	†
Auxiliary verbs	1.389	1.396	1.294				
Common adverbs	**0.884**	**0.905**	**0.837**	*	+		
Conjunctions	**1.098**	**1.084**	**0.926**	*		‡	†
Negations	**0.218**	**0.219**	**0.216**	*			†
Grammar Other							
Regular verbs	**2.833**	**2.841**	**2.645**	*	+		
Adjectives	**0.775**	**0.772**	**0.702**	*		‡	
Comparatives	0.376	0.377	0.333				
Interrogatives	0.188	0.191	0.172				
Numbers	**0.148**	**0.128**	**0.121**	*	+		†
Quantifiers	**0.387**	**0.409**	**0.347**	*	+	‡	
Affect Words	0.796	0.792	0.776				
Positive emotion	0.527	0.537	0.529				
Negative emotion	0.258	0.244	0.236				
Anxiety	0.063	0.061	0.055				
Anger	0.039	0.039	0.035				
Sadness	0.072	0.066	0.063				
Social Words	**1.732**	**1.753**	**1.564**	*		‡	
Family	0.214	0.219	0.186				
Friends	0.074	0.068	0.071				
Female referents	0.291	0.287	0.250				
Male referents	0.307	0.317	0.271				
Cognitive Processes	**1.635**	**1.670**	**1.592**	*	+		†
Insight	**0.349**	**0.349**	**0.347**	*			†
Cause	0.213	0.215	0.186				
Discrepancies	**0.212**	**0.219**	**0.227**	*		‡	†
Tentativeness	**0.323**	**0.339**	**0.315**	*	+		†
Certainty	**0.240**	**0.238**	**0.241**	*			†
Differentiation	**0.411**	**0.428**	**0.392**	*	+		†
Perceptual Processes	0.383	0.377	0.348				
Seeing	0.155	0.158	0.141				
Hearing	0.083	0.080	0.077				
Feeling	0.118	0.111	0.107				
Biological Processes	0.353	0.343	0.314				
Body	0.083	0.073	0.068				
Health/illness	0.158	0.150	0.132				
Sexuality	0.008	0.006	0.007				
Ingesting	0.094	0.101	0.093				
Core Drives and Needs	**1.441**	**1.458**	**1.315**	*		‡	
Affiliation	**0.645**	**0.659**	**0.582**	*			
Achievement	0.247	0.248	0.222				
Power	0.324	0.315	0.287				
Reward focus	0.275	0.279	0.274				
Risk/prevention focus	0.064	0.067	0.063				
Time orientation							
Past focus	**1.486**	**1.422**	**1.289**	*		‡	†
Present focus	**1.034**	**1.080**	**1.072**	*	+		†
Future focus	0.163	0.166	0.165				
Relativity	**2.657**	**2.601**	**2.271**	*		‡	†
Motion	**0.456**	**0.456**	**0.383**	*		‡	†
Space	**1.096**	**1.110**	**0.944**	*	+	‡	†
Time	**1.171**	**1.103**	**0.998**	*		‡	†
Personal Concerns							
Work	0.290	0.288	0.256				
Leisure	0.229	0.231	0.207				
Home	0.134	0.130	0.114				
Money	0.104	0.110	0.093				
Religion	0.020	0.016	0.018				
Death	0.022	0.023	0.017				
Informal Speech	0.051	0.050	0.050				
Swear words	0.005	0.005	0.006				
Netspeak	0.004	0.004	0.004				
Assent	0.014	0.013	0.015				
Nonfluencies	0.023	0.023	0.020				
Fillers	0.003	0.003	0.004				

(a) 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.

Variable	*Major*	*minor*	*noEvent*	*p-values*			
				Mmn	*Mm*	*mn*	*Mn*
Concreteness	**0.333**	**0.338**	**0.311**	*		‡	†
Function words	**8.725**	**7.652**	**8.144**	*		‡	†
Total pronouns	**2.776**	**2.421**	**2.647**	*			†
Personal pronouns	2.102	1.935	1.787				
1st pers singular	1.263	1.084	1.042				
1st pers plural	**0.248**	**0.382**	**0.241**	*	+	‡	
2nd person	0.024	0.006	0.038				
3rd pers singular	0.480	0.362	0.350				
3rd pers plural	0.087	0.101	0.116				†
Impersonal pronouns	**0.673**	**0.486**	**0.860**	*		‡	†
Articles	**1.100**	**1.107**	**0.837**	*		‡	†
Prepositions	**2.231**	**2.056**	**1.785**	*		‡	†
Auxiliary verbs	**1.216**	**0.874**	**1.401**	*	+	‡	†
Common adverbs	**0.779**	**0.615**	**0.887**	*		‡	†
Conjunctions	**0.952**	**0.958**	**0.974**	*			†
Negations	**0.178**	**0.051**	**0.244**	*	+	‡	†
Grammar Other							
Regular verbs	**2.673**	**2.272**	**2.645**	*		‡	†
Adjectives	0.722	0.587	0.715				
Comparatives	0.361	0.239	0.349				
Interrogatives	0.197	0.149	0.168				
Numbers	**0.198**	**0.087**	**0.098**	*	+		†
Quantifiers	**0.298**	**0.343**	**0.342**	*			†
Affect Words	**0.647**	**0.559**	**0.824**	*		‡	†
Positive emotion	**0.408**	**0.458**	**0.601**	*			†
Negative emotion	0.227	0.101	0.213			‡	
Anxiety	0.032	0.031	0.057				†
Anger	0.046	0.011	0.028				
Sadness	0.058	0.042	0.064				
Social Words	**1.816**	**1.702**	**1.462**	*			
Family	**0.330**	**0.230**	**0.178**	*			†
Friends	0.090	0.087	0.054				†
Female referents	**0.398**	**0.256**	**0.228**	*			†
Male referents	0.326	0.261	0.235				
Cognitive Processes	**1.319**	**0.910**	**1.665**	*		‡	†
Insight	**0.293**	**0.197**	**0.339**	*		‡	†
Cause	**0.165**	**0.129**	**0.198**	*			†
Discrepancies	**0.163**	**0.126**	**0.239**	*		‡	†
Tentativeness	**0.235**	**0.222**	**0.321**	*			†
Certainty	**0.162**	**0.132**	**0.254**	*		‡	†
Differentiation	**0.359**	**0.146**	**0.430**	*	+	‡	†
Perceptual Processes	0.386	0.382	0.357				
Seeing	0.180	0.216	0.170				
Hearing	0.102	0.101	0.061				
Feeling	0.089	0.059	0.101				
Biological Processes	**0.416**	**0.385**	**0.285**	*			†
Body	**0.112**	**0.053**	**0.057**	*			†
Health/illness	0.180	0.093	0.125				
Sexuality	**0.019**	**0.000**	**0.006**	*			†
Ingesting	**0.112**	**0.222**	**0.078**	*	+	‡	
Core Drives and Needs	**1.454**	**1.475**	**1.225**	*		‡	
Affiliation	**0.666**	**0.792**	**0.545**	*		‡	
Achievement	0.268	0.174	0.209				
Power	0.334	0.323	0.275				
Reward focus	0.257	0.244	0.255				
Risk/prevention focus	0.055	0.022	0.052				
Time orientation							
Past focus	**1.683**	**1.354**	**1.210**	*			†
Present focus	**0.784**	**0.598**	**1.155**	*		‡	†
Future focus	**0.132**	**0.104**	**0.174**	*			†
Relativity	**2.909**	**2.306**	**2.069**	*			†
Motion	**0.504**	**0.500**	**0.357**	*		‡	†
Space	**1.177**	**1.051**	**0.865**	*		‡	†
Time	**1.292**	**0.812**	**0.905**	*	+		†
Personal Concerns							
Work	0.319	0.219	0.216				
Leisure	**0.237**	**0.348**	**0.180**	*	+	‡	
Home	**0.145**	**0.180**	**0.097**	*		‡	
Money	**0.104**	**0.160**	**0.069**	*		‡	
Religion	0.007	0.008	0.013				
Death	0.031	0.014	0.012				†
Informal Speech	0.045	0.039	0.053				
Swear words	0.010	0.000	0.004				
Netspeak	0.003	0.003	0.009				
Assent	0.011	0.008	0.014				
Nonfluencies	0.020	0.025	0.025				
Fillers	0.001	0.003	0.003				

(b) 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.