Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models

Maarten Sap[‡†*] **Eric Horvitz**[†] **Yejin Choi**[‡◇] **Noah A. Smith**[‡◇] **James W. Pennebaker**[♣]
[†]Microsoft Research
[‡]Paul G. Allen School for Computer Science & Engineering, University of Washington
[◇]Allen Institute for Artificial Intelligence
[♣]Department of Psychology, University of Texas at Austin
msap@cs.washington.edu, horvitz@microsoft.com

Abstract

We investigate the use of NLP as a measure of the cognitive processes involved in storytelling, contrasting *imagination* and *recollection* of events. To facilitate this, we collect and release HIPPOCORPUS, a dataset of 7,000 stories about *imagined* and *recalled* events.

We introduce a measure of *narrative flow* and use this to examine the narratives for imagined and recalled events. Additionally, we measure the differential recruitment of knowledge attributed to *semantic memory* versus *episodic memory* (Tulving, 1972) for *imagined* and *recalled* storytelling by comparing the frequency of descriptions of general *commonsense events* with more specific *realis events*.

Our analyses show that imagined stories have a substantially more linear narrative flow, compared to recalled stories in which adjacent sentences are more disconnected. In addition, while recalled stories rely more on autobiographical events based on episodic memory, imagined stories express more commonsense knowledge based on semantic memory. Finally, our measures reveal the effect of narrativization of memories in stories (e.g., stories about frequently recalled memories flow more linearly; Bartlett, 1932). Our findings highlight the potential of using NLP tools to study the traces of human cognition in language.

1 Introduction

When telling stories, people draw from their own experiences (episodic knowledge; Conway et al., 1996, 2003) and from their general world knowledge (semantic knowledge; Bartlett, 1932; Oatley, 1999). For example, in Figure 1 (top), a recalled story about a birth will likely recount concrete events from that day, relying heavily on the author's episodic memory (Tulving, 1972). On the other hand, an imagined story about a wedding (Figure 1, bottom) will largely draw from the author's commonsense knowledge about the world (Kintsch, 1988; Graesser et al., 1981).

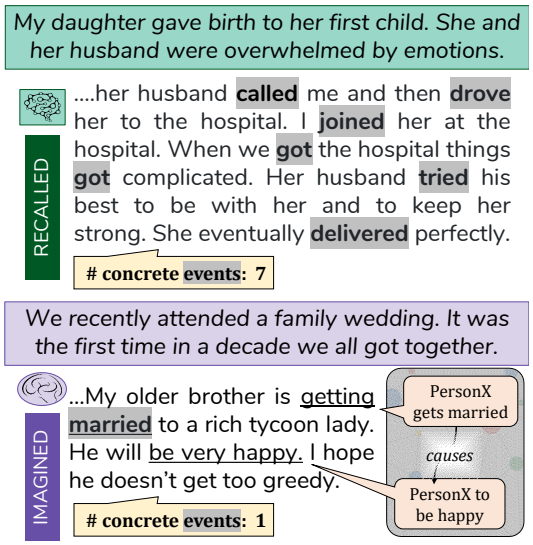

Figure 1: Snippets from two stories from HIPPOCORPUS (top: recalled, bottom: imagined). Concrete or realis events (in gray) are more frequent in recalled stories, whereas general or commonsense events (underlined) are associated with imagined stories.

We harness neural language and commonsense models to study how cognitive processes of recollection and imagination are engaged in storytelling. We rely on two key aspects of stories: *narrative flow* (how the story reads) and *semantic vs. episodic knowledge* (the types of events in the story). We propose as a measure of narrative flow the likelihood of sentences under generative language models conditioned on varying amounts of history. Then, we quantify semantic knowledge by measuring the frequency of commonsense events (from the ATOMIC knowledge graph; Sap et al., 2019), and episodic knowledge by counting realis events (Sims et al., 2019), both shown in Figure 1.

[*] Research conducted during an internship at Microsoft Research.

We introduce HIPPOCORPUS,[1] a dataset of 6,854 diary-like short stories about salient life events, to examine the cognitive processes of remembering and imagining. Using a crowdsourcing pipeline, we collect pairs of recalled and imagined stories written about the same topic. By design, authors of recalled stories rely on their episodic memory to tell their story.

We demonstrate that our measures can uncover differences in imagined and recalled stories in HIPPOCORPUS. Imagined stories contain more commonsense events and elaborations, whereas recalled stories are more dense in concrete events. Additionally, imagined stories flow substantially more linearly than recalled stories. Our findings provide evidence that surface language reflects the differences in cognitive processes used in imagining and remembering.

Additionally, we find that our measures can uncover *narrativization* effects, i.e., the transforming of a memory into a narrative with repeated recall or passing of time (Bartlett, 1932; Reyna and Brainerd, 1995; Christianson, 2014). We find that with increased temporal distance or increased frequency of recollection, recalled stories flow more linearly, express more commonsense knowledge, and are less concrete.

2 HIPPOCORPUS Creation

We construct HIPPOCORPUS, containing 6,854 stories (Table 1), to enable the study of imagined and recalled stories, as most prior corpora are either limited in size or topic (e.g., Greenberg et al., 1996; Ott et al., 2011). See Appendix A for additional details (e.g., worker demographics; §A.2).

2.1 Data Collection

We collect first-person perspective stories in three stages on Amazon Mechanical Turk (MTurk), using a pairing mechanism to account for topical variation between imagined and recalled stories.

Stage 1: recalled. We ask workers to write a 15–25 sentence story about a memorable or salient event that they experienced in the past 6 months. Workers also write a 2–3 sentence summary to be used in subsequent stages, and indicate how long ago the events took place (in weeks or months; TIMESINCEEVENT).

[1] Available at http://aka.ms/hippocorpus.

	# stories	# sents	# words
recalled	2,779	17.8	308.9
imagined	2,756	17.5**	274.2**
retold	1,319	17.3*	296.8**
total	6,854		

Table 1: HIPPOCORPUS data statistics. ** and * indicate significant difference from recalled at $p < 0.001$ and $p < 0.05$, respectively.

Stage 2: imagined. A new set of workers write imagined stories, using a randomly assigned summary from stage 1 as a prompt. Pairing imagined stories with recalled stories allows us to control for variation in the main topic of stories.

Stage 3: retold past. After 2–3 months, we contact workers from stage 1 and ask them to re-tell their stories, providing them with the summary of their story as prompt.

Post-writing questionnaire (all stages). Immediately after writing, workers describe the main topic of the story in a short phrase. We then ask a series of questions regarding personal significance of their story (including frequency of recalling the event: FREQUENCYOFRECALL; see A.1 for questionnaire details). Optionally, workers could report their demographics.[2]

3 Measures

To quantify the traces of imagination and recollection recruited during storytelling, we devise a measure of a story's narrative flow, and of the types of events it contains (concrete vs. general).

3.1 Narrative Flow

Inspired by recent work on discourse modeling (Kang et al., 2019; Nadeem et al., 2019), we use language models to assess the narrative linearity of a story by measuring how sentences relate to their context in the story.

We compare the likelihoods of sentences under two generative models (Figure 2). The *bag* model makes the assumption that every sentence is drawn independently from the main theme of the story (represented by $\mathcal{E}$). On the other hand, the *chain* model assumes that a story begins with a

[2] With IRB approval from the Ethics Advisory Board at Microsoft Research, we restrict workers to the U.S., and ensure they are fairly paid ($7.5–9.5/h).

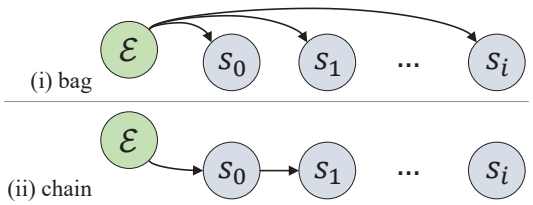

Figure 2: Two probabilistic graphical models representing (i) bag-like and (ii) chain-like (linear) story representations. $\mathcal{E}$ represents the theme of the story.

theme, and sentences linearly follow each other.[3]. Δ_l is computed as the difference in negative log-likelihoods between the bag and chain models:

$$\Delta_l(s_i) = -\frac{1}{|s_i|}[\log p(s_i \mid \mathcal{E}) - \log p(s_i \mid \mathcal{E}, s_{1:i-1})] \quad (1)$$

where the log-probability of a sentence s in a context $\mathcal{C}$ (e.g., topic $\mathcal{E}$ and history $s_{1:i-1}$) is the sum of the log-probabilities of its tokens w_t in context: $\log p(s \mid \mathcal{C}) = \sum_t \log p(w_t \mid \mathcal{C}, w_{0:t-1})$.

We compute the likelihood of sentences using OpenAI's GPT language model (Radford et al., 2018, trained on a large corpus of English fiction), and we set $\mathcal{E}$ to be the summary of the story, but find similar trends using the main event of the story or an empty sequence.

3.2 Episodic vs. Semantic Knowledge

We measure the quantity of episodic and semantic knowledge expressed in stories, as proxies for the differential recruitment of episodic and semantic memory (Tulving, 1972) in stories.

Realis Event Detection We first analyze the prevalence of *realis* events, i.e., factual and non-hypothesized events, such as "I *visited* my mom" (as opposed to irrealis events which have not happened, e.g., "I *should visit* my mom"). By definition, realis events are claimed by the author to have taken place, which makes them more likely to be drawn from from autobiographical or episodic memory in diary-like stories.

We train a realis event tagger (using BERT-base; Devlin et al., 2019) on the annotated literary events corpus by Sims et al. (2019), which slightly outperforms the original author's models. We provide further training details in Appendix B.1.

Semantic and Commonsense Knowledge We measure the amount of commonsense knowledge included explicitly in stories, as a proxy for semantic memory, a form of memory that is thought to encode general knowledge about the world (Tulving, 1972). While this includes facts about how events unfold (i.e., scripts or schemas; Schank and Abelson, 1977; van Kesteren et al., 2012), here we focus on commonsense knowledge, which is also encoded in semantic memory (McRae and Jones, 2013).

Given the social focus of our stories, we use the social commonsense knowledge graph ATOMIC (Sap et al., 2019).[4] For each story, we first match possible ATOMIC events to sentences by selecting events that share noun chunks and verb phrases with sentences (e.g., "getting married" $\rightsquigarrow$ "PersonX gets married"; Figure 1). We then search the matched sentences' surrounding sentences for commonsense inferences (e.g., "be very happy" $\rightsquigarrow$ "happy"; Figure 1). We describe this algorithm in further detail in Appendix B.2. In our analyses, the measure quantifies the number of story sentences with commonsense tuple matches in the two preceding and following sentences.

3.3 Lexical and Stylistic Measures

To supplement our analyses, we compute several coarse-grained lexical counts for each story in HIPPOCORPUS. Such approaches have been used in prior efforts to investigate author mental states, temporal orientation, or counterfactual thinking in language (Tausczik and Pennebaker, 2010; Schwartz et al., 2015; Son et al., 2017).

We count psychologically relevant word categories using the Linguistic Inquiry Word Count (Pennebaker et al., 2015, LIWC;), focusing only on the cognitive processes, positive emotion, negative emotion, and I-word categories, as well as the ANALYTIC and TONE summary variables.[5] Additionally, we measure the average concreteness level of words in stories using the lexicon by Brysbaert et al. (2014).

4 Imagining vs. Remembering

We summarize the differences between imagined and recalled stories in HIPPOCORPUS in Table 2. For our narrative flow and lexicon-based analyses,

[3]Note that this is a sentence-level version of *surprisal* as defined by expectation theory (Hale, 2001; Levy, 2008)

[4]ATOMIC contains social and inferential knowledge about the causes (e.g., "X wants to start a family") and effects (e.g., "X throws a party", "X feels loved") of everyday situations like "PersonX decides to get married".

[5]See liwc.wpengine.com/interpreting-liwc-output/ for more information on LIWC variables.

	measure	effect size (d or β)	direction
	avg. Δ_l (linearity)	0.52^{***}	imagined
	realis events	0.10^{**}	recalled
	commonsense	0.15^{***}	imagined
lexicon-based	ANALYTIC	0.26^{***}	recalled
	concrete	0.13^{***}	recalled
	neg. emo.	0.07^{***}	imagined
	TONE	0.12^{***}	imagined
	I-words	0.17^{***}	imagined
	pos. emo.	0.22^{***}	imagined
	cog. proc.	0.30^{***}	imagined

Table 2: Summary of differences between imagined and recalled stories, according to proposed measures (top), and lexical or word-count measures (bottom). All associations are significant when controlling for multiple comparisons (***: p <0.001; **: p <0.01).

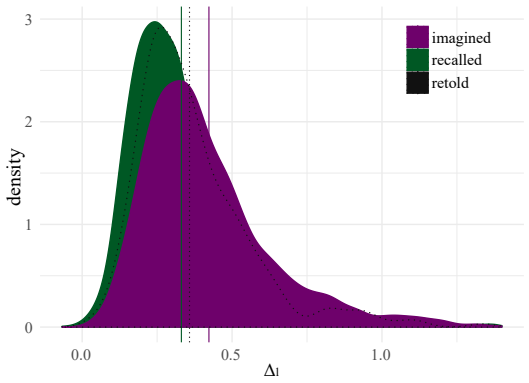

Figure 3: Density plot showing differences in likelihoods of sentences between chain and bag model, for recalled (green), imagined (purple), and retold (dark gray dashed) stories. Vertical lines represent mean Δ_l values for each story type. All three story types differ significantly ($p < 0.001$).

we perform paired t-tests. For realis and commonsense event measures, we perform linear regressions controlling for story length.[6] We Holm-correct for multiple comparisons for all our analyses (Holm, 1979).

Imagined stories flow more linearly. We compare Δ_l, i.e., pairwise differences in NLL for sentences when conditioned on the full history vs. no history (density plot shown in Figure 3). When averaging Δ_l over the entire story, we find that sentences in imagined stories are substantially more predictable based on the context set by prior sentences than sentences in remembered stories. This effect is also present with varying history sizes (see Figure 5 in Appendix C.1).

Recalled stories are more event-dense. As seen in Table 2, we find that imagined stories contain significantly fewer realis events (controlling for story length).[7]

Imagined stories express more commonsense knowledge. Using the same analysis method, our results show that sentences in imagined stories are more likely to have commonsense inferences in their neighborhood compared to recalled stories.

Lexical differences. Lexicon-based counts uncover additional differences between imagined and recalled stories. Namely, imagined stories are more self-focused (I-words), more emotional (TONE, positive and negative emotion) and evoke more cognitive processes.[8] In contrast, recalled stories are more concrete and contain more logical or hierarchical descriptions (ANALYTIC).

Discussion. Our interpretation of these findings is that the consolidated memory of the author's life experience permeates in a more holistic manner through the sentences in the recalled story. Imagined stories are more fluent and contain more commonsense elaborations, which suggests that authors compose a story as a sequence, relying more on preceding sentences and commonsense knowledge to generate the story.

While our findings on linearity hold when using different language models trained on Wikipedia articles (Dai et al., 2019) or English web text (mostly news articles; Radford et al., 2019), a limitation of the findings is that GPT is trained on large corpus of fiction, which may boost linearity scores for imagined (vs. recalled) sentences. Future work could explore the sensitivity of our results to changes in the language model's training domain or neural architecture.

5 Narrativization of Recalled Stories

We further investigate how our narrative and commonsense measures can be used to uncover the narrativization of recalled events (in recalled and retold stories). These analyses aim to investigate the hypothesis that memories are narrativized

[6] Linear regressions use z-scored variables. We confirm that our findings hold with multivariate regressions as well as when adding participant random effects in Appendix C.2.

[7] Note that simply using verb count instead of number of realis events yields the opposite effect, supporting our choice of measure.

[8] The cognitive processes LIWC category counts occurrences of words indicative of cognitive activity (e.g., "think", "because", "know").

over time (Bartlett, 1932), and that distant autobiographical memories are supplemented with semantic or commonsense knowledge (Reyna and Brainerd, 1995; Roediger III et al., 1996; Christianson, 2014; Brigard, 2014).

First, we compare the effects of recency of the event described (TIMESINCEEVENT: a continuous variable representing the log time since the event).[9] Then, we contrast recalled stories to their retold counterparts in pairwise comparisons. Finally, we measure the effect of how frequently the experienced event is thought or talked about (FREQUENCYOFRECALL: a continuous variable ranging from very rarely to very frequently).[10] As in §4, we Holm-correct for multiple comparisons.

Temporal distance. First, we find that recalled and retold stories written about temporally distant events tend to contain more commonsense knowledge ($|\beta| = 1.10$, $p < 0.001$). We found no other significant associations with TIMESINCEEVENT.

On the other hand, the proposed measures uncover differences between the initially recalled and later retold stories that mirror the differences found between recalled and imagined stories (Table 2). Specifically, retold stories flow significantly more linearly than their initial counterparts in a pairwise comparison (Cohen's $|d| = 0.17$, $p < 0.001$; see Figure 3). Our results also indicate that retold stories contain fewer realis events ($|\beta| = 0.09$, $p = 0.025$), and suggest a potential increase in use of commonsense knowledge in the retold stories ($|\beta| = 0.06$, $p = 0.098$).

Using lexicon-based measures, we find that retold stories are significantly higher in scores for cognitive processes ($|d| = 0.12$, $p < 0.001$) and positive tone ($|d| = 0.07$, $p = 0.02$). Surprisingly, initially recalled stories contain more self references than retold stories (I-words; $|d| = 0.10$, $p < 0.001$); higher levels of self reference were found in imagined stories (vs. recalled; Table 2).

Frequency of recall. We find that the more an event is thought or talked about (i.e., higher FREQUENCYOFRECALL), the more linearly its story flows (Δ_l; $|\beta| = 0.07$, $p < 0.001$), and the fewer realis events ($|\beta| = 0.09$, $p < 0.001$) it contains.

Furthermore, using lexicon-based measures, we find that stories with high FREQUENCYOFRECALL tend to contain more self references (I-words; Pearson's $|r| = 0.07$, $p < 0.001$). Conversely, stories that are less frequently recalled are more logical or hierarchical (LIWC's ANALYTIC; Pearson's $|r| = 0.09$, $p < 0.001$) and more concrete (Pearson's $|r| = 0.05$, $p = 0.03$).

Discussion. Our results suggest that the proposed language and commonsense methods can measure the effects of narrativization over time in recalled memories (Bartlett, 1932; Smorti and Fioretti, 2016). On one hand, temporal distance of events is associated with stories containing more commonsense knowledge and having more linear flow. On the other hand, stories about memories that are rarely thought about or talked about are more concrete and contain more realis events, compared to frequently recalled stories which flow more linearly. This suggests that stories that become more narrativized, either by the passing of time or by being recalled repeatedly, become more similar in some ways to imagined stories.

6 Conclusion

To investigate the use of NLP tools for studying the cognitive traces of recollection versus imagination in stories, we collect and release HIPPOCORPUS, a dataset of imagined and recalled stories. We introduce measures to characterize narrative flow and influence of semantic vs. episodic knowledge in stories. We show that imagined stories have a more linear flow and contain more commonsense knowledge, whereas recalled stories are less connected and contain more specific concrete events. Additionally, we show that our measures can uncover the effect in language of narrativization of memories over time. We hope these findings bring attention to the feasibility of employing statistical natural language processing machinery as tools for exploring human cognition.

Acknowledgments

The authors would like to thank the anonymous reviewers, as well as Elizabeth Clark, Tal August, Lucy Lin, Anna Jafarpour, Diana Tamir, Justine Zhang, Saadia Gabriel, and other members of the Microsoft Research and UW teams for their helpful comments.

[9] We use the logarithm of the time elaspsed since the event, as subjects may perceive the passage of time logarithmically (Bruss and Rüschendorf, 2009; Zauberman et al., 2009).

[10] Note that TIMESINCEEVENT and FREQUENCYOFRECALL are somewhat correlated (Pearson $r = 0.05$, $p < 0.001$), and findings for each variable still hold when controlling for the other.

References

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

Felipe De Brigard. 2014. Is memory for remembering? recollection as a form of episodic hypothetical thinking. *Synthese*, 191:155–185.

F. Thomas Bruss and Ludger Rüschendorf. 2009. On the perception of time. *Gerontology*, 56 4:361–70.

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

Sven-Ake Christianson. 2014. *The Handbook of Emotion and Memory: Research and Theory*. Psychology Press.

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).

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):334–340.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

M. Brent Donnellan, Frederick L. Oswald, Brendan M. Baird, and Richard E. Lucas. 2006. The mini-IPIP scales: tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, 18(2):192.

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):1–26.

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):588–602.

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

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Dongyeop Kang, Hiroaki Hayashi, Alan W Black, and Eduard Hovy. 2019. Linguistic versus latent relations for modeling coherent flow in paragraphs. In *EMNLP*.

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).

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

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

Ken McRae and Michael Jones. 2013. Semantic memory. In Daniel Reisberg, editor, *The Oxford Handbook of Cognitive Psychology*, Psychology Publications.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with Discourse-Aware neural models. In *Workshop on Innovative Use of NLP for Educational Applications @ ACL*.

Keith Oatley. 1999. Why fiction may be twice as true as fact: Fiction as cognitive and emotional simulation. *Review of general psychology*, 3(2):101–117.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. Linguistic inquiry and word count: LIWC 2015.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Unpublished.

Valerie F. Reyna and Charles J. Brainerd. 1995. Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, 7(1):1–75.

Henry L Roediger III, J Derek Jacoby, and Kathleen B McDermott. 1996. Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language*, 35(2):300–318.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1:1–20.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

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

H. Andrew Schwartz, Gregory Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Jonah Berger, Martin Seligman, and Lyle Ungar. 2015. Extracting human temporal orientation from Facebook language. In *NAACL*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *ACL*.

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):296–319.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *ACL*.

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):24–54.

Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological review*, 117(2):440.

Endel Tulving. 1972. Episodic and semantic memory. *Organization of Memory*, 1:381–403.

Endel Tulving and Daniel L. Schacter. 1990. Priming and human memory systems. *Science*, 247(4940):301–306.

Gal Zauberman, B Kyu Kim, Selin A Malkoc, and James R Bettman. 2009. Discounting time and time discounting: Subjective time perception and intertemporal preferences. *Journal of Marketing Research*, 46(4):543–556.

(a) Recalled main events

(b) Imagined main events

Figure 4: We extract phrases from the main themes of recalled (left) and imagined (right) stories, using RAKE (Rose et al., 2010); size of words corresponds to frequency in corpus, and color is only for readability.

A Data Collection

We describe the data collection in further detail, and release our MTurk annotation templates.[11]

A.1 Post-Writing Questionnaire

After each writing stage (recalled, imagined, retold), we ask workers to rate "how impactful, important, or personal" the story was to them (for imagined and recalled stories), "how similar" to their own lives the story felt (imagined only), and "how often [they] think or talk about the events" in the story (recalled only), on a Likert scale from 1–5. Workers also take the four "openness" items from the Mini-IPIP personality questionnaire (Donnellan et al., 2006) as an assessment of overall creativity. Finally, workers optionally report their demographic information (age, gender, race).

A.2 Worker Demographics

Our stories are written by 5,387 unique U.S.-based workers, who were 47% male and 52% female (<1% non-binary, <1% other). Workers were 36 years old on average (s.d. 10 years), and predominantly white (73%, with 10% Black, 6% Hispanic, 5% Asian). We find no significant differences in demographics between the authors of imagined and recalled stories,[12] but authors of imagined stories scored slightly higher on measures of creativity and openness to experience (Cohen's $d = 0.08$, $p = 0.01$).

Note that we randomly paired story summaries to workers. We did not attempt to match the demographics of the recalled story to the demographics of the imagined author. Future work should investigate whether there are linguistic effects of differing demographics between the two authors.[13]

B Episodic vs. Semantic Knowledge

B.1 Realis Events

To detect realis events in our stories, we train a tagger (using BERT-base; Devlin et al., 2019) on the annotated corpus by Sims et al. (2019). This corpus contains 8k realis events annotated by experts in sentences drawn from 100 English books. With development and test F_1 scores of 83.7% and 75.8%, respectively, our event tagger slightly outperforms the best performing model in Sims et al. (2019), which reached 73.9% F_1. In our analyses, we use our tagger to detect the number of realis event mentions.

B.2 Commonsense Knowledge Matching

We quantify the prevalence of commonsense knowledge in stories, as a proxy for measuring the traces of semantic memory (Tulving and Schacter, 1990). Semantic memory is thought to encode commonsense as well as general semantic knowledge (McRae and Jones, 2013).

We design a commonsense extraction tool that aligns sentences in stories with commonsense tuples, using a heuristic matching algorithm. Given a story, we match possible ATOMIC events to sentences by selecting events that share noun chunks and verb phrases with sentences. For every sentence s_i that matches an event E in ATOMIC, we check surrounding sentences for mentions of commonsense inferences (using the same noun and verb phrase matching strategy); specifically, we

[11] Available at http://aka.ms/hippocorpus.

[12] We run Chi-squared tests for gender ($\chi^2 = 1.01$, $p = 0.80$), for age ($\chi^2 = 9.99$, $p = 0.26$), and for race ($\chi^2 = 9.99$,$p = 0.35$).

[13] Future work could investigate social distance alongside other types of psychological distances (e.g., physical, temporal), using the framework given by Construal Theory (Trope and Liberman, 2010).

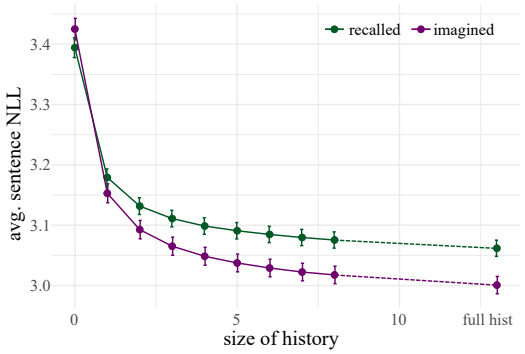

Figure 5: Average negative log likelihood (NLL) of sentences conditioned on varying sizes of histories of included sentences for recalled (green) and imagined (purple) stories (with 95% confidence intervals). For history sizes > 1, differences are significant when controlling for multiple comparisons ($p < 0.001$).

check the n_c preceding sentences for matches of causes of E, and the n_e following sentences for event E's effects.

To measure the prevalence of semantic memory in a story, we count the number of sentences that matched ATOMIC knowledge tuples in their surrounding context. We use a context window of size $n_c = n_e = 2$ to match inferences, and use the spaCy pipeline (Honnibal and Montani, 2017) to extract noun and verb phrases.

C Recollection vs. Imagination

C.1 Linearity with Varying Context Size

Shown in Figure 5, we compare the negative log-likelihood of sentences when conditioned on varying history sizes (using the story summary as context $\mathcal{E}$). As expected, conditioning on longer histories increases the predictability of a sentence. However, this effect is significantly larger for imagined stories, which suggests that imagined stories flow more linearly than recalled stories.

variable	β w/o rand. eff.	β w/ rand. eff.
story length	0.319***	0.159**
Δ_l (linearity)	-0.454***	-0.642***
realis events	0.147***	0.228***
commonsense	-0.144***	-0.157***

Table 3: Results of multivariate linear regression models (with and without participants random effects), regressing onto story type (0: imagined vs. 1: recalled) as the dependent variable. All effects are significant (**: $p < 0.005$, ***: $p < 0.001$).

C.2 Robustness of Findings

To confirm the validity of our measures, we report partial correlations between each of our measures, controlling for story length. We find that our realis measure is negatively correlated with our commonsense measures (Pearson $r = -0.137$, $p < 0.001$), and positively correlated with our linearity measure ($r = 0.111$, $p < 0.001$). Linearity and commonsense were not significantly correlated ($r = -0.02$, $p = 0.21$).

Additionally, we confirm that our findings still hold when controlling for other measures and participant random effects. Notably, we find stronger associations between our measures and story type when controlling for other measures, as shown in Table 3. We see a similar trend when additionally controlling for individual variation in workers.