Trends in Cognitive Sciences

CellPress

Opinion

Studying memory narratives with natural language processing

Can Fenerci[1], Ziming Cheng[2,3], Donna Rose Addis[2,3,4], Buddhika Bellana[2,5], and Signy Sheldon[1,*]

Cognitive neuroscience research has begun to use natural language processing (NLP) to examine memory narratives with the hopes of gaining a nuanced understanding of the mechanisms underlying differences in memory recall, both across groups and tasks. However, the diversity of NLP approaches can make it challenging for researchers to know which techniques to use and when to apply them. We outline how different NLP techniques can be applied to narrative descriptions to address specific questions about the neurocognitive processes underlying memory narratives. We also discuss the strengths and limitations of NLP methods for use in memory research, highlighting both their potential and their constraints in uncovering the mechanisms of remembering.

Natural language processing in memory research

Cognitive neuroscience research has increasingly prioritized understanding cognition through a lens of ecological validity [1,2]. In the study of **episodic memory** (see Glossary), this focus has led researchers away from using highly controlled memoranda (e.g., lists of unrelated words) toward stimuli that more closely approximate real-world experiences (e.g., narratives) as testbeds for memory research [3–8]. This movement toward studying **memory narratives** and memories as narratives [9] presents its own challenges, namely, how does one go about quantifying open-ended recall of complex, interrelated events? Traditionally, researchers overcome this problem using manual scoring methods to segment transcripts of the recalled narratives (e.g., autobiographical memories) into discrete events or details and then calculate the number of these recalled details [10–16]. However, these manual scoring techniques are labor-intensive, requiring substantial time and expertise to train the scorers, and rely on the scorer's interpretation of the described event and scoring rubric. Moreover, many of these scoring methods have been developed to characterize the quantity of detail within a narration, or **memory specificity** [10], which may not be best suited to capture other theoretically important dimensions of memory expression (e.g., sentiment, sequencing of events, or recurrent themes).

To capture these dimensions, researchers are increasingly turning to **NLP**. NLP is a field rooted in computer science, linguistics, and artificial intelligence that aims to build machines that can recognize, understand, and produce natural human language. In recent years, the field has undergone rapid progress, resulting in a deluge of readily available computational tools for efficiently representing and analyzing text data. These techniques range from basic representation of word co-occurrence statistics in a corpus [17] to sophisticated modern **large language models** (**LLMs**; e.g., ChatGPT). This rapid surge and diversity of NLP tools, while exciting, pose practical challenges for researchers outside computational linguistics who may be uncertain of how best to apply these methods and how to deal with the various decision points to be made when using these tools (see Box 1). Here, we highlight how NLP has and can be used to address critical and unanswered questions about the neurocognitive processes of episodic memory in the context

Highlights

How memories are narrated can reveal insights into the underlying neurocognitive processes and psychological states of an individual.

Traditionally, memory narratives are manually scored to quantify memory specificity, neglecting other theoretically important dimensions of memory expression (e.g., consistency, emotionality, themes, etc.).

Natural language processing (NLP) offers new, automatic, and scalable ways for memory researchers to gain these insights.

NLP methods can describe how we encode, store, and recall past personal memories and, when complemented by neuroimaging data, can reveal the underlying neurocognitive processes of remembering.

However, NLP methods come with their own set of limitations, the understanding of which can mitigate against bias while improving the interpretability of results.

[1]Department of Psychology, McGill University, Montreal, Québec, Canada
[2]Rotman Research Institute, Baycrest Academy for Research and Education, Toronto, Ontario, Canada
[3]Department of Psychology, University of Toronto, Toronto, Ontario, Canada
[4]School of Psychology, The University of Auckland, Auckland, New Zealand
[5]Department of Psychology, Glendon Campus, York University, Toronto, Ontario, Canada

*Correspondence:
signy.sheldon@mcgill.ca (S. Sheldon).

Box 1. NLP analytic decision points to consider

Methodological decisions during NLP analysis play a significant role in shaping the outcome of the analysis. Many of these decision points arise when (i) preprocessing the raw text data and (ii) selecting the language model to use for conducting the analysis.

(i) Preprocessing text data. There are different preprocessing steps used to clean the raw text data before analysis, in which a researcher must decide if they will exclude certain words or special characters from the text. For example, a researcher must decide if and what types of words to exclude from the text. Should punctuation and/or stop words devoid of explicit semantic meaning (e.g., 'and', 'or', 'the', etc.) be excluded from the text? The stop words are often removed as they can artificially inflate the NLP measures (e.g., similarity among texts), given their ubiquity in everyday language. However, stop words may hold some value unbeknownst to a researcher, especially if the model being used handles stop words contextually (e.g., BERT and GPT), in which case removing them may disrupt sentence meaning. Should temporal markers or causal connectors (e.g., 'then', 'because', 'after', etc.) be removed from the text? The decision to remove these words can improve clarity and reduce dimensionality of large text datasets. Removing these words also benefit models, such as bag-of-words models, by focusing the output on the content words, yielding more interpretable results. Nonetheless, removing them can obscure the narrative flow of the text, making it harder to detect how memories are organized.

(ii) Selecting the language model. After preprocessing, a researcher must select the models to use to conduct their analysis. A key distinction is between models pretrained on existing data corpus from different sources (blogs, forums, and movie reviews) via an unsupervised learning approach versus self-trained models. The pretrained models offer significant advantages; they are ready to use, saving time and computational resources. However, pretrained models may not capture nuances specific to a given research question. This is because pretrained models are representative only of their training data (e.g., language models trained on text produced by younger adults will likely not accurately capture features of text produced by older adults). If a researcher decides that a pretrained model is not sufficient or appropriate to answer their research question, self-trained models can be constructed through an extant corpus. While self-trained models offer better alignment with a given dataset than pretrained models, there is a risk of inadequate training data. Without sufficient quantity or diversity of input, these models may produce narrow or inaccurate representations of memory narratives. A compromise between these models is via fine-tuning of pretrained models. Fine-tuning enables a model to adapt (tune) a pretrained model to the specific context, with the downside of being significantly technical and requiring substantial resource investment.

In sum, while there is no one correct way to make these decisions when conducting NLP analysis, an awareness of the tradeoffs associated with these decision points can allow researchers to mitigate biases and improve the interpretability of their results.

of narrative memory (for discussions on the role of NLP in cognitive science more broadly, see [18–21]). To this end, we focus on how three classes of NLP (and related) tools (linguistic feature analysis, **text vectorization**, and **topic modeling)** can address important questions about narrating our personal experiences from memory (see Table 1 and Figure 1 for a summary of these tools).

Linguistic features: what can words tell us about memory narratives?

One of the earliest forms of NLP involved extracting **linguistic features** from text or speech to understand the cognitive processes underlying narrations (Figure 1, top). The precise linguistic feature of focus will depend on what the researcher hopes to understand about the psychological or cognitive status of the narrator. Some common features that have been extracted to meet this aim are word usage [22], such as the use of affective or emotional words [23], parts of speech (e.g., nouns, verbs, adjectives), and sentence structure (e.g., subject-verb-object), which index language complexity or richness [24].

For example, in a landmark study, researchers analyzed the linguistic complexity of the diary entries of over 600 nuns, assessing features such as the use of diverse parts of speech and the number of distinct ideas or pieces of information conveyed per unit of text [24]. Nuns whose writings showed greater linguistic complexity were less likely to develop Alzheimer's disease or experience cognitive decline in older age. That is, linguistic complexity in early adulthood is positively associated with cognitive function later in life, illustrating more generally the use of linguistic markers to predict cognitive status.

Glossary

Data corpus: a large collection of text documents (written or spoken) that is used to analyze language patterns or train LLMs for natural language processing.

Embeddings: numerical representations of text as vectors in a continuous, high-dimensional space. In the context of NLP, embeddings can capture the semantic relationships between words, phrases, or entire documents.

Episodic memory: detailed memories of specific experiences and events from one's life that include details about when and where these events occurred. Episodic memory is often contrasted against semantic memory, which is memory for general knowledge and facts.

Large language models (LLMs): a category of machine learning models designed to understand and generate natural language. These models are often created through training on large amounts of data (books, internet sites) to understand human language.

Linguistic features: characteristics of language. Most relevant to memory narratives are the syntactic features (arrangement of words and phrases), lexical features (i.e., vocabulary and words choices), semantic features (i.e., meaning of words or sentences), and discourse features (i.e., cohesion or coherence)

Memory narratives: verbal descriptions of a past event that are organized and interpreted as a cohesive story to covey a particular experience.

Memory specificity: the degree to which a memory narrative includes specific details, particularly those related to spatial and temporal context details, that allow a vivid reconstruction of a past event. Memory specificity depends on episodic memory and has been linked to cognitive deficits in aging and disorders like depression or post-traumatic stress disorder.

Natural language processing (NLP): a host of computational methods for automatically analyzing text. NLP techniques enable the analysis, understanding, and generation of natural language, allowing for the processing of large volumes of text data to extract meaningful patterns and insights.

Text vectorization: the processing of converting unstructured text into a structured numerical format that is

Table 1. Research questions related to the reviewed NLP approaches[a]

NLP method	Potential research question(s)	Example case use	Refs
Linguistic feature analysis	What is the affect or emotional tone of a narrated event and how does this measurement relate to the type of event being narrated (e.g., traumatic event) and the person narrating the event?	Using LIWC, Kredlow *et al.* showed that the amount of negative language in narratives collected 2 weeks after 9/11 predicted memory accuracy up to 10 years later	[23]
	What type of language is used within a narration (e.g., complexity, inclusion of certain parts of speech)? What is the extent to which the use of this language reveals insights into the cognitive processes underlying the narration?	Using VADER, Rouhani *et al.* showed that memories from the COVID-19 pandemic were more negative than other months within the same year	[32]
Text vectorization and embeddings	How accurately is an encoded event narrated from memory? Does this change across tasks and groups?	Fenerci et al. used USE and showed greater similarity between older adults' recollections across different retrieval goals	[40]
	How consistently is a memory recalled across retrieval scenarios (e.g., shifting goals, audiences), groups (younger versus older adults), or elapsed time?	Using BERT, Fowler *et al*. showed higher similarity for coimagined events than for individually imagined events	[42]
Topic modeling	What is the underlying meaning conveyed within a narrated event?	Sheldon *et al*. used LDA and showed that older adults include a greater diversity of topics in autobiographical memories than younger adults	[47]
	What is the common theme presented in a corpus of narrated events among a group? Does this theme reveal shared meaning and ideas?	Yeung *et al*. showed thematic differences between positive and negative involuntary autobiographical memories of younger adults using LDA	[45]

[a]This table presents a brief description of possible research questions related to narrative research associated with the three reviewed natural language processing analytic approaches, presented with example cases from published research (refer to the text for a more thorough description of these cases).

suitable for processing in NLP computational models.

Topic modeling: a statistical technique used in NLP to estimate latent or hidden themes or topics within a dataset that are often represented as words that co-occur together in the dataset.

More contemporary approaches to linguistic feature analysis are lexicon-based methods, which leverage predefined lexicons to categorize words from narrations into psychological dimensions (e.g., concreteness, emotion, tone) and output a summary score for each dimension (e.g., average concreteness [25]). A popular lexicon-based method for text analysis is linguistic inquiry and word count (LIWC [26]). LIWC uses over 100 predefined lexicons to score a given document in terms of its usage of basic parts of speech (e.g., pronouns, nouns, verbs) or how it refers to more abstract psychological features (e.g., social referents, emotion, past focused), providing a useful approach to identify characteristics of language use between scenarios and across groups. In a recent example, researchers examined the memory narrations of patients who confabulate (i.e., produce false memories) and individuals who do not confabulate and compared the types of speech used in these narrations. The result was that confabulating patients were more likely to use informal speech in their recollections than healthy controls [27], suggesting that this type of speech could be used as a marker of different states of interest (e.g., confabulation) in future cases.

Studies have used linguistic feature extraction from narrations to delineate individuals with and without mental health conditions and, as a result, help us understand certain conditions. One study used LIWC to analyze memories from individuals with major depressive disorder (MDD) [28]. Results showed that those with MDD used more self-referential pronouns when recounting negative memories and fewer self-referential pronouns when recounting positive memories than

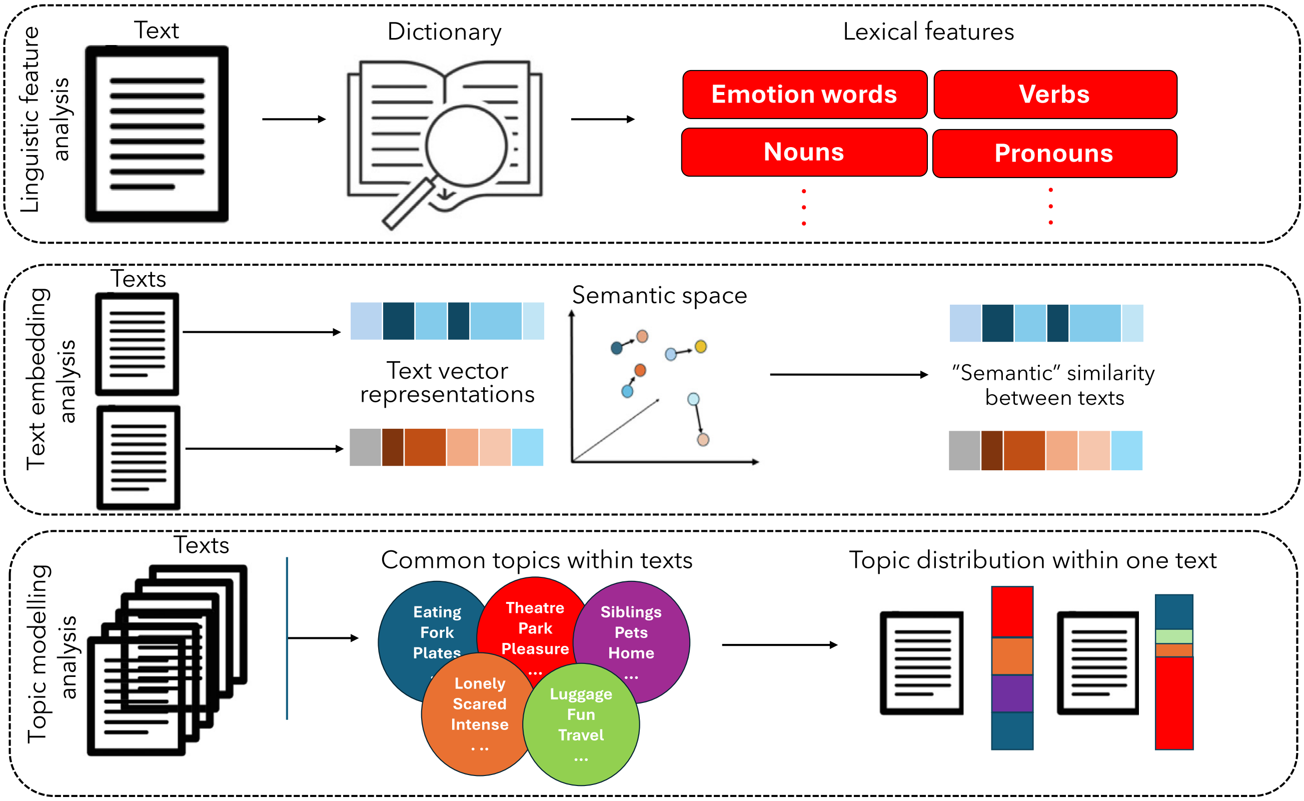
Linguistic feature analysis
Text
Dictionary
Lexical features
Emotion words
Verbs
Nouns
Pronouns
Text embedding analysis
Texts
Text vector representations
Semantic space
"Semantic" similarity between texts
Topic modelling analysis
Texts
Common topics within texts
Eating Fork Plates ...
Theatre Park Pleasure ...
Siblings Pets Home ...
Lonely Scared Intense ...
Luggage Fun Travel ...
Topic distribution within one text
Trends in Cognitive Sciences

Figure 1. Summary of three classes of natural language processing (NLP) tools. A visualization of the three reviewed NLP analytic approaches for studying memory that are reviewed in this article and can address important questions about how we narrate our past personal experiences from memory. The top panel depicts linguistic feature analysis, which leverages predefined dictionaries (i.e., lexicons) to categorize words from narratives to psychological dimensions based on the characteristics of language (e.g., semantic or lexical features, syntax, etc.), The middle panel depicts text vectorization/embedding analysis, which converts a given text (e.g., word, sentence, or documents) into numerical vectors in high-dimensional semantic space. Researchers can then compute the similarity between these vectors to approximate the similarity in meaning between them. Finally, the bottom panel depicts topic modeling, which identifies common topics within texts and returns a list of words and their relative contributions.

healthy controls, reflecting the heightened negative focus inherent to MDD. Similarly, examining the linguistic patterns of recollections of traumatic events such as the 9/11 attacks revealed that individuals who used fewer first-person pronouns when describing the events also exhibited fewer post-traumatic stress disorder symptoms [29,30]. Following upon this work, a recent analysis of over 600 memory narratives about 9/11 events found that the amount of negative language in narratives collected 2 weeks after the event predicted memory accuracy up to 10 years later [23]. This result underscores how the precise language used in memory narratives can predict later recall accuracy and aligns with broader research linking linguistic features to cognitive and mental health.

Earlier lexicon-based methods like LIWC require an explicit mapping between words and their meaning as well as word-counting algorithms to measure linguistic features, which can prove challenging given the productivity of human languages. Modern lexical feature approaches to NLP are LLMs that use data-driven strategies, deriving the latent associations between words and meaning by modeling their co-occurrence in massive online **data corpus**. Other lexicon-based approaches are those that incorporate rules to enhance the flexibility of their lexicons. For example, Valence Aware Dictionary and sEntiment Reasoner (VADER [31]) tracks word order-sensitive relationships between terms to compute the emotional tone of a text (i.e., positive to negative valence). Instead of relying exclusively on a lexicon with a fixed mapping between a word and valence, VADER also tracks punctuation (e.g., !) and intensifiers (e.g., extremely, somewhat, kind of), affording it additional sensitivity to the degree of the sentiment being expressed. In a recent study, VADER was used to quantify the sentiment of autobiographical memories collected in March 2020, during the global coronavirus disease 2019 (COVID-19) pandemic [32]. Memories from this period were more negative than other months within the same year, and more negative sentiment was associated with greater recall for the corresponding month [33].

Text vectorization and embeddings: how do memory narratives change?

The above-noted methodological developments in NLP have also been critical for researchers to gain insights into higher-level representations and meaning within and across memories. While these examples demonstrate the utility of lexicon-based approaches in probing human memory, they remain limited, particularly for the research questions that focus on changes in meaning beyond the words used to narrate an experience. In this section, we review the use of text vectorization via language embedding models to compare the content of a memory narration across time and scenarios and between individuals (Figure 1, middle). In brief, embedding models convert a given text (e.g., word, sentence, or document) into numerical vectors, thus 'embedding' the text vectors in a high-dimensional semantic space. These 'spaces' are derived from large corpora of natural text (e.g., entirety of Wikipedia) and represent semantics by inferring what a word means based on how it was used in the training corpus. The specific ways in which **embeddings** are computed vary considerably, from directly modeling word occurrence statistics (e.g., Global Vectors for Word Representation (GLoVe) [34]) to training neural networks to complete a specific task, like predicting the text that is likely to appear before or after a specific target [e.g., word2vec [35] and Universal Sentence Encoder (USE) [36]]. Yet, contemporary embedding models leverage the powerful transformer architecture (e.g., Bidirectional Encoder Representations from Transformers (BERT)), which has given rise to the proliferation of LLMs that we see today [e.g., Generative Pretrained Transformer (GPT)].

Although embeddings may differ in how they are computed across models, memory researchers can use these models to vectorize memory narratives and compute the similarity between these vectors as an approximation for the similarity in meaning between them. Depending on what these text vectors represent, researchers can quantify narrative memory accuracy and consistency, and these embeddings can even be used to predict memory for naturalistic events.

For example, if a memory researcher has access to the description of an encoded experience that participants are asked to recall, embeddings can be used to estimate memory accuracy. Imagine that participants read a short story and are asked to freely recall that same story after a delay. In this case, the researcher can estimate memory accuracy by comparing the similarity between vectors representing the encoded material (i.e., short story) and participants' recollections of this material, which quantifies the extent to which the encoded and retrieved content correspond to one another. For example, researchers have used embedding models, like the USE [35], to measure the content similarity between annotations of encoded video clips and participants' recollections of those clips to capture content reinstatement during recall [37]. Embedding an encoded experience in semantic space has also proven useful in exploring more implicit instances of memory, like persistence of a recent experience in an individual's spontaneous thoughts. For example, word embeddings were used to show that the content of a free word generation task remains semantically related to the core themes of a recently read story for several minutes after reading it [38].

Embeddings can also provide insights into changes in how consistent recalled memories are across different remembering contexts (e.g., shifting goals, audiences) and across participant groups (e.g., younger vs. older adults) [40,41,53], both known to affect how memories are recollected [10,39]. This approach does not require having an encoding event but instead requires multiple narrations of the same event across the factors of interest (e.g., autobiographical memories encoded outside the lab) [53]. We used this approach to estimate the content overlap between narrations of a short movie when retrieval goals differed (i.e., emphasizing accuracy vs. social reasons for recall) and by younger and older adults [40]. Comparing the text vectors representing recollections of the same event across these two distinct retrieval goals and age groups, we found that there was greater similarity between recollections in older adults than in younger adults, suggesting an age-related deficit in flexibly shifting retrieval content to meet distinct retrieval goals. Using a similar approach, researchers have shown that participants who show similar neural responses to salient changes in a movie (i.e., event boundaries) also showed similarity in how they recollect the movie [41], suggesting that these changes shape how narratives are appraised by individuals. By providing a way to quantify consistency across memories, text vectorization and embedding models have opened many new and exciting avenues for memory research, such as characterizing how individuals coimagine future events [42] and determining how common it is for a person to change their story when recounting an event.

Beyond using embeddings to characterize the memory itself, embeddings have also been used to predict the aspects of a complex experience most likely to be remembered. For example, USE was used to estimate the network structure of stories [6]. Briefly, participants segmented a series of short films into their constituent events, and sentence embeddings were computed based on event-specific annotations. Using the event-specific embeddings, the researchers were able to estimate the semantic relatedness between all events within a film, resulting in a ‘narrative network’. Events that were more semantically related to other events (i.e., higher semantic centrality) were more likely to be recalled in a separate group of participants.

Topic modeling: what is the underlying meaning of the memory narrative?

Topic modeling aims to derive latent themes in a corpus that are composed of words that tend to co-occur with one another (Figure 1, bottom). Prior to NLP, researchers have relied on qualitative thematic analytic approaches to characterize topics in a corpus, which is known to be very time-consuming and highly subjective (although valuable for particular research questions [43]). NLP-based topic modeling provides an alternate and data-driven way to estimate topics in open-ended memory data.

Embedding approaches, as described above, can be used to extract semantic content (topics) by estimating a text vector’s relative position to other embeddings within a given space, but more common topic modeling approaches are probabilistic topic models, such as latent Dirichlet allocation (LDA) [44]. These models implement machine learning to derive latent psychological constructs via probability distributions and identify hidden topics within a curated data corpus, returning a list of words and their relative contributions to (i.e., topic weight) a set of higher-order topics. These topics can then be applied to a memory narrative to determine the distribution of topics (not just one topic as manual coding) within that memory narration, which can then be compared across situations. This approach can be particularly useful for memory scientists interested in the content of individual narratives rather than how multiple narratives compare to one another, for which embedding approaches are helpful.

Topic modeling can be used to answer two critical questions within the field of memory, the first being how the meaning of a memory changes over time. For instance, topic modeling was used on memories collected from a large sample of Americans in 2020 [32]. After identifying the dominant topics presented in these memories, they then examined how these topics were distributed across memory recollections across the year. While personal topics, like work or social events, were evenly distributed throughout the months, collective topics, such as those representing public events like the COVID-19 pandemic lockdown and the presidential election, were most prevalent around the time of their occurrence in the public arena, showing how collective experiences shape our autobiographical memories.

Second, topic modeling can answer questions about how the meaning of a memory differs as a function of what is being remembered and who is remembering it. For example, researchers used LDA to study thematic differences between positive and negative involuntary autobiographical memories in a group of younger adults. Across participants, these memories contained fundamentally different topics (e.g., ‘arguments’ in negative versus ‘vacations’ in positive memories) [45]. In a follow-up study, the topics of recurrent involuntary memories collected over 2 years were extracted from a large cohort of participants and related to symptoms of mental health disorders. The results showed that the extracted topics selectively predicted symptom levels of certain mental health disorders, suggesting that what meaning is occurring in these involuntary memories is disorder specific and could reflect current concerns of the affected individual [46].

As with the above case example, researchers using topic modeling will often try to understand the concepts expressed within the extracted topics, which are groups of words. As a result, researchers have to evaluate these groupings and assign a label to each topic for interpretation, which could introduce human biases and errors. For example, if a researcher is looking for the presence of the topic ‘family’ within a memory dataset, they might label a topic dictionary that has the words ‘sister, love, anger, house, hammer’ as a ‘family’ topic, yet these words could plausibly be described using another term such as ‘renovations’.

One way to circumvent this evaluation step is to focus on topic distributions rather than evaluate the topics per se, both within memories and across memories [47]. For example, we used LDA to examine the diversity of topics used in narrated autobiographical memories of younger and older adults to understand how the breath of meanings within a memory differs by the age of the narrator. To this end, we calculated the topic distributions within personal memories narrated by younger and older adults and then compared the maximum topic distribution score between these age groups. We found that older adults showed a lower maximum topic score, indicative of including a greater diversity of topics than younger adults [47]. Another example of using topic modeling without labelling the extracted topics comes from a study in which the researchers used LDA to extract topic vectors from text annotations describing the temporal unfolding of a television episode (BBC’s Sherlock) and examined how these topics, or mini events, were present in the narrated recollections of this movie by a group of participants. The result was that participants commonly remembered topics that described the general narrative but not specific details, showcasing commonalities in the general event meaning that cut across participants.

Linking NLP analysis to brain activity: insights into the neural basis of memory

NLP methods on their own provide a window into how individuals remember; however, these methods are increasingly paired with neuroimaging methods to further our understanding of how the brain encodes and reconstructs narrative memories [6,48,49]. For example, in one study, the embedding approach was used to model not only memory performance but also the changes in patterns of brain activity across individuals [6]. Events that were more semantically related to other events in the encoded movie clips were associated with greater activity and more similar neural responses across individuals during recall, particularly in regions of the default mode network (DMN). This finding suggests that the semantic relatedness across events in a narrative (derived using NLP-based embedding approaches) is an important feature of how we remember complex, naturalistic events.

Topic modeling approaches have similarly been paired with brain imaging data to reveal underlying neural mechanisms. For instance, outside of memory research, topic modeling has been used to decode how the brain represents linguistic meaning [50]. Within the field, a recent fMRI study scanned participants as they watched and later recalled an audiovisual movie [48]. Researchers used topic modeling to identify shared topics in participants' recall and linked these to brain activity, finding that the posterior–medial subsystem of the DMN (involved in forming internal representations) was sensitive to the organization of topics in a memory. As such, these studies illustrate the potential of combining NLP with neuroimaging methods to understand not only the core brain systems involved in forming and expressing memory narrations but also the mechanisms that drive differences in how we narrate memories (see Outstanding questions).

Concluding remarks

NLP, a host of computational methods for automatically analyzing text, is transforming how researchers can and have been studying memory (see Box 2). Together, these methods offer powerful ways to answer questions about how individuals encode, store, and recall personal memories in ways not accessible or scalable through manual scoring of memories. NLP methods have the potential to reveal hidden structures in text data not apparent to human coders or not feasible to manually quantify in large corpora. When combined with data on the rememberer's cognitive status or complemented by neuroimaging data, these methods can map the underlying neurocognitive processes of remembering, revealing new insights into human memory and potentially even predict the trajectory of an individual's brain health (see Outstanding questions).

However, it is important to understand that there are limits to what these methods can tell us about the attributes of memories. First, although NLP methods may seem objective, with pipelines offering increased reproducibility, replicability, and reliability over manual coding, they are not entirely without bias (see Box 1). From the corpora used to train a model, preprocessing steps, and data analysis applied to the outputs, there are several decision points that can impact the results to varying degrees. Moreover, the direct interpretation of the outputs can be difficult or impossible, such as labelling extracted topics or interpreting dimensions in an embedding space, respectively. Furthermore, not all memory datasets are appropriate for each NLP approach. For

Box 2. Other applications of NLP methods for memory science

NLP offers ways to answer novel questions about how and what we remember, yet they can also chart new ways to refine existing methods. First, NLP methods are currently being used to streamline and automatize status quo methods for scoring details within narratives. For example, new work has emerged to automatically score memory narratives to quantify the number of specific events [54–56] or event-specific details remembered by participants [57,58] that is traditionally the work of manual scorers [10]. For example, an LLM (i.e., distilBERT) was fine-tuned with previously scored free recall data from autobiographical memory and future imagination tasks [57]. The researchers found a strong relationship between the model predicted and human-scored number of details within the narratives, indicating promise for reducing the cost of manual detail scoring. Second, work has begun to use NLP methods to estimate how narratives are structured [59,60]. For example, GPT-3 was used to quantify narrative flow in thousands of autobiographical and imagined events based on the extent to which sentences flow from their preceding context [59]. The results showed greater narrative flow in imagined compared with autobiographical narratives, offering novel ways to rapidly and reliably analyze narratives. Additionally, the same LLM (i.e., GPT-3) was used to identify when one event ends and another begins (i.e., event boundaries) in a continuous narrative [60]. Typically, event boundaries are approximated by averaging across human annotations of these boundaries with variable agreement between these annotators. The researchers found that GPT-3-derived annotations are significantly correlated with human annotations of event boundaries, suggesting that these annotations can reliably and objectively pinpoint the precise location of these boundaries. Together, these tools break away from traditional approaches to score memory narratives and hold promise in enabling researcher to scale up their studies with larger sample sizes and datasets than typically used.

Outstanding questions

Given the increasing parallels between NLP and human cognition, will there come a point when these measurements are equivalent? This question follows the understanding that NLP methods can provide different measurements of memory narratives than those based on manual, human hand scoring or personal reports of memories.

Can NLP approaches be used to reveal how deviations or modifications to memory occur (e.g., retrieval errors, intrusions, omissions, etc.)? In their current state, NLP approaches can tell us the extent to which memory recall deviates from the encoded event but not the nature of these deviations.

What are the ethical implications of using NLP to analyze autobiographical memories? Given that autobiographical memories can contain sensitive content that can be personally identifying, what privacy or consent considerations need to be addressed when analyzing these data?

What design considerations should be taken into account for experiments using NLP approaches?

How can we evaluate the efficacy of NLP approaches in measuring what we think they are measuring? Not all NLP models and approaches are equal, particularly when considering the many decisions to be made when building or using these models, such as the dictionary a language model is built on or how the model is fine-tuned (see Box 1).

What individual differences can reliably be determined by NLP methods from memory narratives? There are several known important individual difference factors that affect how a person narrates a memory, ranging from imagery to emotional regulation ability.

If NLP methods can provide insights into an individual's cognitive status, can they also be useful in predicting the risk of developing memory-related disorders (e.g., dementia, mental health conditions) and in informing therapeutic interventions?

instance, training a topic model on a limited set of memories or shorter narratives can lead to unreliable or unstable estimation of topics.

More broadly, NLP methods, at least in the current state, are not as effective as humans at capturing the nuanced ways culture and contextual factors, such as social norms, motivational states, and intended audience, can affect memory narratives, nor nonliteral (e.g., gestural) and expressive elements of natural language use (e.g., prosody) that accompany language [51,52]. To this point, we consider that the human examination of narrations, embracing the associated subjectivity, brings a unique perspective to memory research that cannot be replicated by NLP (but see Outstanding questions). Thus, while NLP methods are exciting ways to study narrated memories, they are best considered one of many tools in a memory researcher's toolbox (e.g., hand scoring and manual thematic analysis).

Acknowledgments

This research was supported by funding from the Natural Sciences and Engineering Research Council of Canada and Canada Research Chair Program awards to S.S.

Declaration of interests

We report that there are no competing interests to declare.

References

1. Nastase, S.A. *et al.* (2020) Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage* 222, 117254
2. Sonkusare, S. *et al.* (2019) Naturalistic stimuli in neuroscience: critically acclaimed. *Trends Cogn. Sci.* 23, 699–714
3. Addis, D.R. *et al.* (2007) Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia* 45, 1363–1377
4. Fenerci, C. *et al.* (2024) Shift happens: aging alters the content but not the organization of memory for complex events. *Aging Neuropsychol. Cogn.* 32, 118–141
5. Lee, H. *et al.* (2020) What can narratives tell us about the neural bases of human memory? *Curr. Opin. Behav. Sci.* 32, 111–119
6. Lee, H. and Chen, J. (2022) Predicting memory from the network structure of naturalistic events. *Nat. Commun.* 13, 4235
7. Baldassano, C. *et al.* (2017) Discovering event structure in continuous narrative perception and memory. *Neuron* 95, 709–721
8. Scheurich, R. *et al.* (2021) Evidence for a visual bias when recalling complex narratives. *PLoS One* 16, e0249950
9. Palombo, D.J. (2024) Beyond memory: the transcendence of episodic narratives. *Can. J. Exp. Psychol.* 78, 155–162
10. Levine, B. *et al.* (2002) Aging and autobiographical memory: dissociating episodic from semantic retrieval. *Psychol. Aging* 17, 677–689
11. Simpson, S. *et al.* (2023) Effects of healthy and neuropathological aging on autobiographical memory: a meta-analysis of studies using the autobiographical interview. *J. Gerontol. Ser. B* 78, 1617–1624
12. Renoult, L. *et al.* (2020) Classification of general and personal semantic details in the autobiographical interview. *Neuropsychologia* 144, 107501
13. Kian, T. *et al.* (2021) Tell me about your visit with the lions: eliciting event narratives to examine children's memory and learning during summer camp at a local zoo. *Front. Psychol.* 12, 657454
14. Van Abbema, D. and Bauer, P. (2005) Autobiographical memory in middle childhood: recollections of the recent and distant past. *Memory* 13, 829–845
15. Thorndyke, P.W. (1977) Cognitive structures in comprehension and memory of narrative discourse. *Cogn. Psychol.* 9, 77–110
16. Mandler, J.M. and Johnson, N.S. (1977) Remembrance of things parsed: story structure and recall. *Cogn. Psychol.* 9, 111–151
17. Salton, G. (1991) Developments in automatic text retrieval. *Science* 253, 974–980
18. Demszky, D. *et al.* (2023) Using large language models in psychology. *Nat. Rev. Psychol.* 2, 688–701
19. Frank, M.C. (2023) Openly accessible LLMs can help us to understand human cognition. *Nat. Hum. Behav.* 7, 1825–1827
20. Jackson, J.C. *et al.* (2022) From text to thought: how analyzing language can advance psychological science. *Perspect. Psychol. Sci.* 17, 805–826
21. Ke, L. *et al.* (2024) Exploring the frontiers of LLMs in psychological applications: a comprehensive review. *arXiv*, Published online January 3, 2024. https://doi.org/10.48550/arXiv.2401.01519
22. Pasupathi, M. (2007) Telling and the remembered self: linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory* 15, 258–270
23. Kredlow, M.A. *et al.* (2024) Emotion language use in narratives of the 9/11 attacks predicts long-term memory. *Emotion* 24, 808–819
24. Snowdon, D.A. *et al.* (1996) Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. *JAMA* 275, 528–532
25. Brysbaert, M. *et al.* (2014) Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res.* 46, 904–911
26. Boyd, R.L. (2022) The Development and Psychometric Properties of LIWC-22. *ResearchGate*, Published online February 1, 2022. https://doi.org/10.13140/RG.2.2.23890.43205
27. Balshin-Rosenberg, F. *et al.* (2024) It's not a lie…If you believe it: narrative analysis of autobiographical memories reveals overconfidence disposition in patients who confabulate. *Cortex* 175, 66–80
28. Himmelstein, P. *et al.* (2018) Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PLoS One* 13, e0207814
29. D'Andrea, W. *et al.* (2012) Linguistic predictors of post-traumatic stress disorder symptoms following 11 September 2001. *Appl. Cogn. Psychol.* 26, 316–323
30. Dekel, S. and Bonanno, G.A. (2013) Changes in trauma memory and patterns of posttraumatic stress. *Psychol. Trauma Theory Res. Pract. Policy* 5, 26–34
31. Hutto, C. and Gilbert, E. (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc. Int. AAAI Conf. Web Soc. Media* 8, 216–225
32. Rouhani, N. and Stanley, D. (2023) COVID-Dynamic Team & Adolphs, R. Collective events and individual affect shape autobiographical memory. *Proc. Natl. Acad. Sci.* 120, e2221919120

33. Martin, C.B. *et al.* (2022) A smartphone intervention that enhances real-world memory and promotes differentiation of hippocampal activity in older adults. *Proc. Natl. Acad. Sci.* 119, e2214285119
34. Pennington, J., *et al.* GloVe: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Moschitti, A., *et al*., eds), pp. 1532–1543, Association for Computational Linguistics
35. Mikolov, T. *et al.* (2013) Efficient estimation of word representations in vector space. *arXiv*, Published online January 16, 2013. https://doi.org/10.48550/arXiv.1301.3781
36. Cer, D. *et al.* (2018) Universal sentence encoder. *arXiv*, Published online March 29, 2018. http://arxiv.org/abs/1803.11175
37. Shen, X. *et al.* (2023) Machine-learning as a validated tool to characterize individual differences in free recall of naturalistic events. *Psychon. Bull. Rev.* 30, 308–316
38. Bellana, B. *et al.* (2022) Narrative thinking lingers in spontaneous thought. *Nat. Commun.* 13, 4585
39. Dutemple, E. and Sheldon, S. (2022) The effect of retrieval goals on the content recalled from complex narratives. *Mem. Cogn.* 50, 397–406
40. Fenerci, C. *et al.* (2024) The impact of retrieval goals on memory for complex events in younger and older adults. *SSRN*, Published online May 15, 2024. https://doi.org/10.2139/ssrn.4829205
41. Sava-Segal, C. *et al.* (2023) Individual differences in neural event segmentation of continuous experiences. *Cereb. Cortex* 33, 8164–8178
42. Fowler, Z. *et al.* (2024) Collaborative imagination synchronizes representations of the future and fosters social connection in the present. *Proc. Natl. Acad. Sci.* 121, e2318292121
43. Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101
44. Blei, D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022
45. Yeung, R.C. *et al.* (2022) Understanding autobiographical memory content using computational text analysis. *Memory* 30, 1267–1287
46. Yeung, R.C. and Fernandes, M.A. (2023) Specific topics, specific symptoms: linking the content of recurrent involuntary memories to mental health using computational text analysis. *NPJ Mental Health Res.* 2, 22
47. Sheldon, S. *et al.* (2023) Differences in the content and coherence of autobiographical memories between younger and older adults: insights from text analysis. *Psychol. Aging* 39, 59–71
48. Heusser, A.C. *et al.* (2021) Geometric models reveal behavioural and neural signatures of transforming experiences into memories. *Nat. Hum. Behav.* 5, 905–919
49. Zada, Z. *et al.* (2024) A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron* 112, 3211–3222
50. Pereira, F. *et al.* (2018) Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* 9, 963
51. Berger, J. and Packard, G. (2022) Using natural language processing to understand people and culture. *Am. Psychol.* 77, 525–537
52. Băroiu, A.-C. and Trăuşan-Matu, Ş. (2022) Automatic sarcasm detection: systematic literature review. *Information* 13, 399
53. Addis, D. *et al.* (2025) How depression, age and specificity shapes the content of autobiographical thoughts. *OSF Preprints*, Published online at January 20. https://doi.org/10.31219/osf.io/fbhg8
54. Mistica, M. *et al.* (2024) A natural language model to automate scoring of autobiographical memories. *Behav. Res. Methods* 56, 6707–6720
55. Takano, K. *et al.* (2018) Computerized scoring algorithms for the Autobiographical Memory Test. *Psychol. Assess.* 30, 259–273
56. Takano, K. *et al.* (2017) Unraveling the linguistic nature of specific autobiographical memories using a computerized classification algorithm. *Behav. Res. Methods* 49, 835–852
57. van Genugten, R.D. and Schacter, D.L. (2024) Automated scoring of the autobiographical interview with natural language processing. *Behav. Res. Methods* 56, 2243–2259
58. Peters, J. *et al.* (2017) Quantitative text feature analysis of autobiographical interview data: prediction of episodic details, semantic details and temporal discounting. *Sci. Rep.* 7, 14989
59. Sap, M. *et al.* (2022) Quantifying the narrative flow of imagined versus autobiographical stories. *Proc. Natl. Acad. Sci.* 119, e2211715119
60. Michelmann, S. *et al.* (2023) Large language models can segment narrative events similarly to humans. *arXiv*, Published online January 24, 2023. https://doi.org/10.48550/arXiv.2301.10297