A Bayesian Model of Memory for Text

Mark Andrews (mark.andrews@ntu.ac.uk)
Department of Psychology, Nottingham Trent University
Nottingham, NG1 4FQ, UK

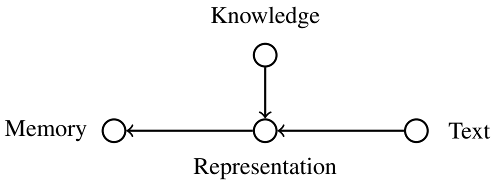
Abstract

The study of memory for texts has had an long tradition of research in psychology. According to most general accounts of text memory, the recognition or recall of items in a text is based on querying a memory representation that is built up on the basis of background knowledge. The objective of this paper is to describe and thoroughly test a Bayesian model of this general account. In particular, we develop a model that describes how we use our background knowledge to form memories as a process of Bayesian inference of the statistical patterns that are inherent in a text, followed by posterior predictive inference of the words that are typical of those inferred patterns. This provides us with precise predictions about what words will be remembered, whether veridically or erroneously, from any given text. We then test these predictions using data from a memory experiment using a relatively large sample of randomly chosen texts from a representative corpus of British English.

Keywords: Bayesian models; Memory; Reconstructive memory; Text memory;

Introduction

The seminal study on memory for text[1] is usually attributed to Bartlett (1932). In this now classic work, Bartlett argued that a person's memory for what they read is based on a reconstruction of the information in the text that is strongly dependent on their background knowledge and experiences. From this seminal work, but especially since the widespread adoption of schema based accounts of text memory beginning in the 1970's (e.g., Mandler & Johnson, 1977; Schank & Abelson, 1977; Bower, Black, & Turner, 1979), there has been something close to a consensus on the broad or general characteristics of human text memory. According to this general account — which we can summarize by the following schematic:



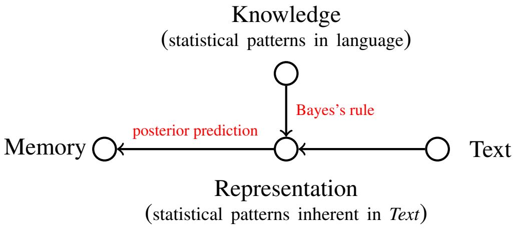
— the recognition or recall of items in a text is based on querying a representation of the text that is built up on the basis of background knowledge and experience.

Although some variant of this general account is widely held, it is essentially an informal and untestable theory. Certainly, there has been ample evidence showing that we use our background knowledge to make inferences and associations concerning text content and that these inferences then influence our memory (e.g. Bransford, Barclay, & Franks, 1972; Graesser, Singer, & Trabasso, 1994; Zwaan & Radvansky, 1998; Rawson & Kintsch, 2002, to name but a few). However, in most studies, even fundamental concepts such as memory schemas are not formally defined (see, e.g. Ghosh & Gilboa, 2014), and ostensibly formal models of knowledge influences on text representation, such as the well known work of Kintsch (1988), often require hand-coding of background knowledge and text structures and can only be applied to small and contrived examples. Consequently, there is no formal or computational account of how background knowledge is used to infer a representation of text content and how memories are then derived from this representation that is sufficiently precise to lead to testable empirical predictions.

In this paper, following general principles followed by Hemmer and Steyvers (2009a, 2009b, 2009c) in their studies on memory for visual objects and natural scenes, we describe a probabilistic model that uses Bayesian inference to infer a representation of a text's content on the basis of background knowledge and then uses posterior predictive inference to represent the memories of that text. This provides us with precise predictions about what words will be remembered, whether veridically or erroneously, from any given text. We then test these predictions using data from a memory experiment using a relatively large sample of randomly chosen texts from a representative corpus of British English.

Probabilistic Model

We begin with the assumption that our background knowledge that is relevant for our memory of text is primarily knowledge of the statistical patterns across spoken and written language. Given any probabilistic language model that specifies these statistical patterns, as we explain below, we may then use Bayes's rule to infer which patterns are inherent in any given text. From this, we may then predict, via posterior predictive inference, which words are and are not typical or compatible with the inferred statistical representation of the text. This effectively serves as the memory of the content of the text. As such, this provides a computational description of the previous schematic, i.e.,


Knowledge
(statistical patterns in language)
Bayes's rule
posterior prediction
Memory
Text
Representation
(statistical patterns inherent in *Text*)

In practical terms, we have many options for our choice of probabilistic language model. However, *probabilistic topic*

[1] In this paper, we use the term *text* to refer generally to any coherent or self-contained piece of spoken or written language.

models (see, e.g. Griffiths, Steyvers, & Tenenbaum, 2007; Steyvers & Griffiths, 2007; Blei, 2012) have proved highly effective in capturing the statistical patterns that characterize the coarse-grained "discourse topics" across spoken and written language. Here, we use a type of probabilistic topic model known as a *hierarchical Dirichlet process mixture model* (HDPMM) (Teh, Jordan, Beal, & Blei, 2006).

A HDPMM is a probabilistic generative model of bag-of-words[2] language data. It treats a corpus of language data as a set of J texts $w_1, w_2 \ldots w_j \ldots w_J$, where text j, i.e., w_j, is a set of n_j words from a finite vocabulary, represented simply by the V integers $\{1, 2 \ldots V\}$. From this, we have each w_j defined as $w_j = w_{j1}, w_{j2} \ldots w_{ji} \ldots w_{jn_j}$, with each $w_{ji} \in \{1 \ldots V\}$. As a generative model of this corpus, the HDPMM treats each observed word w_{ji} as a sample from one of an underlying set of component distributions, $\phi_1, \phi_2 \ldots \phi_k \ldots$, where each ϕ_k is a probability distribution over $\{1 \ldots V\}$. Each ϕ_k effectively identifies a "discourse topic". For example, here is a sample of 6 topics from an inferred model, where we show the 7 most probable words in each topic:

theatre	music	league	prison	rate	pub
stage	band	cup	years	cent	guinness
arts	rock	season	sentence	inflation	beer
play	song	team	jail	recession	drink
dance	record	game	home	recovery	bar
opera	pop	match	prisoner	economy	drinking
cast	dance	division	serving	cut	alcohol

The identity of the particular topic distribution from which w_{ji} is drawn is determined by the value of a discrete latent variable $x_{ji} \in \{1, 2 \ldots k \ldots\}$ that corresponds to w_{ji}. The probability distribution over the possible values of each x_{ji} is given by a categorical distribution π_j, i.e., $\pi_j = \pi_{j1}, \pi_{j2} \ldots \pi_{jk} \ldots$, where $0 \leq \pi_{jk} \leq 1$ and $\sum_{k=1}^{\infty} \pi_{jk} = 1$, that is specific to text j. Each π_j is assumed to be drawn from a Dirichlet process prior whose base distribution, m, is a categorical distribution over the positive integers and whose scalar concentration parameter is a. The m base distribution is assumed to be drawn from a stick breaking distribution with a parameter γ. As such, the generative model of the corpus is as follows:

$$\begin{aligned} w_{ji}|x_{ji}, \phi \sim \mathrm{dcat}(\phi_{x_{ji}}), \quad & x_{ji}|\pi_j \sim \mathrm{dcat}(\pi_j), & i \in 1 \ldots n_j \\ & \pi_j|a, m \sim \mathrm{ddp}(a, m), & j \in 1 \ldots J \\ & m|\gamma \sim \mathrm{dstick}(\gamma), & \end{aligned}$$

where dcat is a categorical probability distribution, ddp is a Dirichlet process, and dstick is a stick breaking distribution. The prior on the component distributions $\phi_1 \ldots \phi_k \ldots$ was a Dirichlet distribution with concentration parameter b and length V location parameter ψ.

Having inferred a HDPMM on the basis of a corpus of language data $\mathcal{D}$, given any new text, $w_{j'}$, we can use Bayes's rule to infer the posterior probability over $\pi_{j'}$, which is the probability distribution over the discourse topics in $w_{j'}$:

$$\mathrm{P}(\pi_{j'}|w_{j'}, \mathcal{D}) \propto \mathrm{P}(w_{j'}|\pi_{j'}, \mathcal{D})\mathrm{P}(\pi_{j'}|\mathcal{D}).$$

We may then use the posterior predictive distribution to infer the words that are typical of the topics inherent in $w_{j'}$. The predicted probability of word $w_{j'i'}$ given text $w_{j'}$ is given by

$$\mathrm{P}(w_{j'i'}|w_{j'}, \mathcal{D}) = \int \mathrm{P}(w_{j'i'}|\pi_{j'}, \mathcal{D})\mathrm{P}(\pi_{j'}|w_{j'}, \mathcal{D})d\pi_{j'}$$

Corpus As our language corpus, we used the British National Corpus (BNC) (BNC Consortium, 2007). From the entire BNC, we extracted all sections that were tagged as paragraphs. This gave us a corpus with a total word count of 87,564,696 words. From this, we created a set of 184,271 texts, each between 250 and 500 words long. These were created by using either single paragraphs in this count range, or concatenating consecutive paragraphs until they were within this range. The total word count of this set of texts was 78,723,408 words. We then restricted the word types by excluding words that occurred less than 5 times in total, and any words on either of two lists of stopwords, and any words that were not listed in a dictionary of $\approx$ 60K English words. This lead to a final vocabulary of 49,328 word types. For more information, see Footnote[3].

Inference We used a Gibbs sampler to infer the posterior distribution over the values of latent variables, i.e., $\{x_{ji} \colon j \in 1 \ldots J, i \in 1 \ldots n_j\}$, as well as the hyper-parameters m, a, b, ψ, and γ. For more information, see Footnote[4]

Prediction From the entire set of paragraphs in the BNC, we randomly sampled 50 paragraphs whose length was 150 ± 10 words, where at least 90% of the words are in the aforementioned dictionary of English words, and where at least 75% of the words were in a set of words for which word association norms exists (see the following section for more details on the word association norms we used). For more information, see Footnote[5].

For each of the 50 sampled texts, we then used posterior predictive inference, as described above, to obtain the probability distribution over words that are typical or compatible with the topic based representation of each text. As explained above, this distribution effectively provides the inferred model's memory of the content of the text. A Gibbs sampler was used to infer each text's posterior distribution over π, which is the probability distribution over discourse topics in that text. Two example texts and their posterior predictive inferences are shown in Figure 1. For more information, see Footnote[6].

[2] According to a bag-of-words model, a language corpus is a set of texts, where each text is an unordered set, or bag, of words.

[3] Full details about how the corpus was created, including all the code used to create it, is available at `https://github.com/lawsofthought/tantalum`

[4] Full details about the Gibbs sampler for the HDPMM, including the code implementing it, can be found at `https://lawsofthought.github.io/gustavproject`.

[5] Full details about how we sampled the texts, including the code implementing the sampling and the sampled texts themselves, can be found at `https://github.com/lawsofthought/berkelium`.

[6] Full details about how we sampled from the posterior predictive distribution, including the code implementing the sampling, can be found at `https://github.com/lawsofthought/gallium`.

Improve your mood and counteract stress: Ask anyone who exercises regularly and they will tell you that they always feel exhilarated at the end of a session even if they had begun by feeling that they were not in the mood for exercise and had almost forced themselves to continue. Physical fitness also provides considerable protection against stress and the illnesses it can cause. So, however busy your life, perhaps you could try and fit some regular exercise into your day. Let it be something which is in complete contrast to the way you normally spend your time. One word of warning though: if you are someone whose daily life involves a strong competitive element, you would do well to avoid too much in the way of competitive sport (squash, tennis and so on) as your form of exercise as these will only tend to maintain an already high level of stress.

****

Developmental norms are an attempt to provide an indication of the ages at which one might expect ordinary children to show evidence of certain skills or abilities. Since children vary with respect to the ages at which they demonstrate any particular behaviour, norms represent an average obtained from an examination of the developmental changes occurring in a large number of children. Data from a large sample will show the earliest age at which a child would be expected to gain control of a particular aspect of language, and the age by which 90 per cent or 95 per cent of non-handicapped children might be expected to show evidence of the same ability. If children who have already been diagnosed as suffering from some specific handicapping condition are included, the data will show the expected age delay before this group matches the performance of the normally developing children.

****

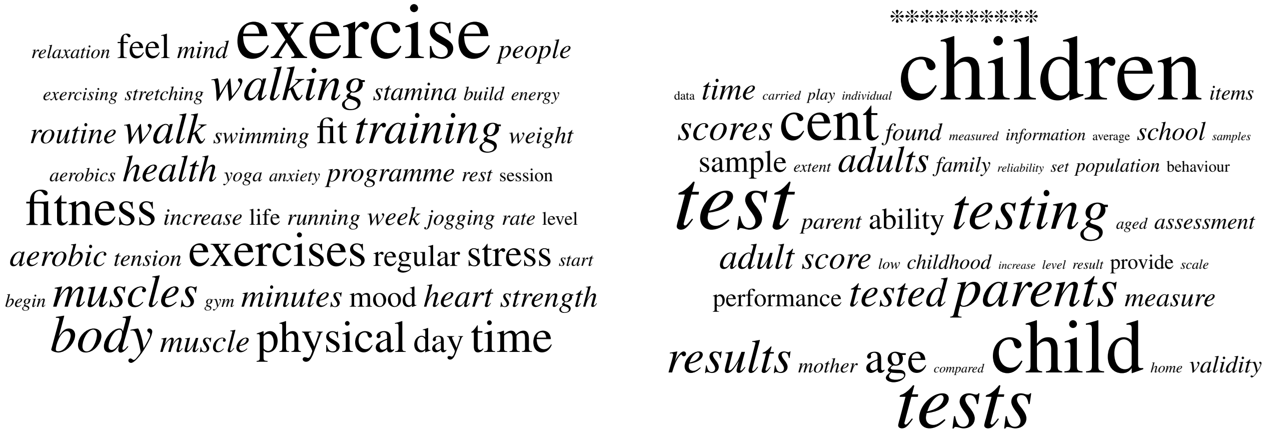


Figure 1: An example of two of the texts used in the memory experiment, and samples from the HDPMM's posterior prediction for each one. The predicted words are scaled as a function of their predicted probability, and we show the 50 most highly predicted words (excluding stopwords and words not in the vocabulary) for each text. Words in italics are predicted words that were not in the text itself. These, in effect, are the model's false memories.

Comparison models

The focus of our analysis is whether the probability of recognizing or recalling any given word having read a particular text is predicted by our HDPMM's posterior predictive distribution over words for that text. To properly evaluate the model's predictions, it is necessary to compare them to those of other plausible models. Here, we will compare the Bayesian model to predictions made by two *associative* models. Both of these models predict that the words that are remembered from a text are those that are most associated, on average, with the text's content. Associative models are strong models to compare to the Bayesian model because associative strength has been repeatedly shown to a strong predictor of memory for words in word lists (e.g., Roediger, Watson, McDermott, & Gallo, 2001; Gallo, 2006).

The statistical co-occurrence probability of two words, w_k and w_l, which we will denote $\mathrm{P_c}(w_k, w_l)$, is defined as the empirical probability of observing word w_k and w_l in the same text[7] in the language. Here, we calculate $\mathrm{P_c}(w_k, w_l)$ using the same BNC corpus as was used above, i.e. with the same 184,271 texts each between 250 and 500 words. From this, we can calculate

$$\mathrm{P_c}(w_k|w_l) = \frac{\mathrm{P_c}(w_k, w_l)}{\mathrm{P_c}(w_l)},$$

which is the conditional probability of observing w_k in any text given that w_l has been observed. From this, if $\mathrm{text}_j = w_{j1}, w_{j2} \ldots w_{jn_j}$, the predicted association probability of word w_k according to text_j is

$$\mathrm{P_c}(w_k|\mathrm{text}_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathrm{P_c}(w_k|w_{ji}).$$

We can interpret this value intuitively as the average association between w_k and text_j, with association defined in terms of statistical co-occurrences in the language.

An alternative means to calculate the average association between w_k and text_j is using word association norms, rather

[7] Here, as above, we use the term *text* to denote any coherent and self-contained piece of language.

than statistical co-occurrences. If A_{kl} is the frequency that word w_k is stated as associated with word w_l, then the conditional probability of word w_k given w_l is

$$\mathrm{P}_{\mathrm{A}}(w_k|w_l) = \frac{A_{kl}}{\sum_{i=1}^{V} A_{il}},$$

where V is the total number of words in our vocabulary of response words. Now, given $\text{text}_j = w_{j1}, w_{j2} \ldots w_{jn_j}$, we can calculate

$$\mathrm{P}_{\mathrm{A}}(w_k|\text{text}_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathrm{P}_{\mathrm{A}}(w_k|w_{ji}),$$

which we can interpret as the average association between w_k and text_j, with association now defined in terms of word association norms rather than statistical co-occurrences. Though a large set of English word association norms are available from the widely used Nelson norms (Nelson, McEvoy, & Schreiber, 2004), we used an even larger set that is a pre-release of the English *small world of words* association norms (De Deyne & Storms, 2017). This provided word associates, produced by 101,119 participants, to 10,050 word types. For more information, see Footnote[8].

Experiment

Our aim in this experiment is to measure participants' memory of the 50 sampled texts described in the previous section. Participants read these texts at their normal reading speed and then their memory for what they have read is tested using both recall and recognition tasks. We will then compare the pattern of results from our participants with the predictions of the models.

Methods

Participants 216 people (113 female, 103 male) participated in the experiment. The ages ranged from 17 to 78 years, with a median of 34 years. Participants were recruited from the student and general populations, with the only restriction being that they be native English speakers.

Design Pre-experiment sample size determination calculations showed that, given the reasonable assumptions of both inter-text and inter-subject variability in memory performance, a relatively large number of texts and participants was necessary. In particular, we showed that there is a high probability of detecting effects, even when these effects are relatively weak, if we have at least 50 texts and at least 150 subjects are used. Importantly, these results hold even when each subject sees only a small subset of total number of texts, and this subset can be as low as 3 texts per each participant. We therefore used all 50 texts described above, and initially aimed for approximately 200 participants, with each participant being tested with a randomly sample of 3 texts.

Materials The texts used as stimuli for this experiment were the above mentioned 50 texts.

For the recognition tasks, test word lists with 20 words each were created. Of the 20 words in each list, 10 were present in the to-be-memorized text, while the remaining 10 were not present in it. For each text, the list was created as follows. Key words were extracted from each text and also from the surrounding paragraphs to that text in the BNC. This was done by calculating the tfidf (term frequency, inverse document frequency) value for each word, and then applying a threshold to exclude the less informative words. 10 words were then randomly selected from the key words of each text. A further 10 words were randomly sampled from the key words of the surrounding paragraphs excluding any words the in the main text itself. This set of 10 words were therefore not present in the text to be memorized, but given that they were selected from surrounding paragraphs, they were likely to be meaningfully related to it. As such, they would serve a useful items on the recognition memory test as they could not easily be dismissed without a proper search of memory. For more information, see Footnote[9].

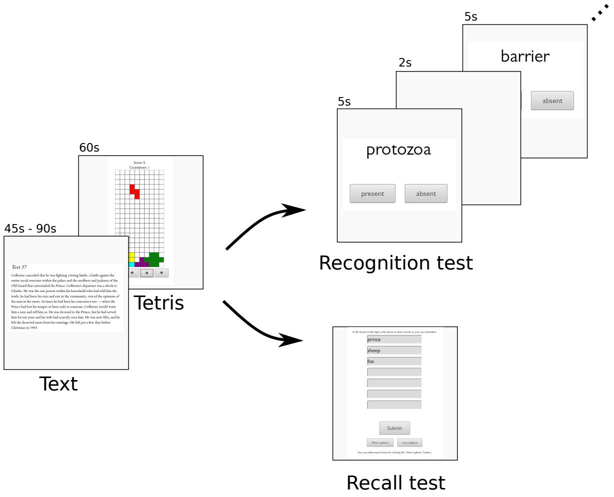


Figure 2: The task diagram of one block in the experiment: Participants read a randomly assigned text, perform a filler task, and then have their memory tested using either a recognition or recall test, with the test type being randomly chosen. This process is repeated three times for each participant.

Procedure Each experiment session proceeded as follows (see also Figure 2):

- After initial information and instructions, which informed participants that they would be engaging in memory tasks, one of the sample texts appeared on screen. Participants were instructed to read this text at their normal reading. The text stayed on screen for a maximum of 90 seconds,

[8] Full details about how these two associative models were created, including the code implementing them, can be found at `https://github.com/lawsofthought/gallium`.

[9] Full details about the recognition test word lists were created, including the code implementing this, can be found at `https://github.com/lawsofthought/berkelium`.

but after 45 seconds, participants were able to move on the next screen if they so wished.

- On the following screen participants were asked to play the computer game *Tetris* for exactly 60 seconds.
- At the completion of the game, participants proceeded to the memory task. For each participant and for each text, the memory test was randomly chosen to be either a recognition or a recall task.
  - For the recognition test, the 20 test items were presented on screen, one word at a time, with an inter-stimulus-interval of 2 seconds. They remained on screen for 5 seconds or until the subject indicated with a button press whether the word shown was *present* or *absent* from the text. No feedback was given after each response.
  - If the participant was assigned to the recall test, a screen of a list of small empty text boxes was presented where and they were asked to type as many words as they could remember, one word into each text box. Initially, 10 empty texts boxes were presented, and more boxes could be added with a button press.
- Upon completion of the memory test, participants were given the option of pausing or proceeding to the next test. Each participant performed three tests in total, with the three texts to which they were assigned being always randomly sampled from the set of 50 texts.

The experiments were presented using the *Wilhelm*[10] web-browser based experiment presentation software that was hosted at *https://www.cognitionexperiments.org*. This software allowed the experiment to be done any web-browser based device, e.g., phones, tablets, laptops and desktops.

Results

For more information about the results, see Footnote[11].

Descriptives In the recognition memory tests trials, the overall accuracy rate was 76%. Overall, the false positive rate, i.e. where participants responded "present" to words that were not present in the text they read, was 27%. The false negative rate, i.e. where participants responded "absent" to words that actually were present in the text, was 22%. For the recall tests, the median number of recalled words per each test was 7, with between 2 and 15 words recalled in 95% of tests. The overall accuracy of recall was 70%, and thus there was an overall false recall rate of 30%.

Model evaluation For the recognition memory data, we model how well each model predicts the behavioural results using a random effects logistic regression model. In other words, for each of the models being evaluated, we fit the recognition memory data using the same random effects logistic regression but using a different predictor variable in each case. The logistic regression model is

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \alpha_{s_i} + \alpha_{t_i} + (\beta_{s_i} + \beta_{t_i} + \beta)\phi_i + bx_i,$$

where i indexes the experiment trial, p_i is the probability of the participant responding "present" to the word presented on trial i, s_i is the identity of the participant on trial i, t_i is the identity of the text on trial i, ϕ_i is the log of the model's predicted probability of the word on trial i, x_i indicates if the word on trial i was present in text t_i. The random effects regression coefficients are α_{s_i}, α_{t_i}, β_{s_i}, β_{t_i}, which are modelled as drawn from zero-mean Normal distributions.

Having fit the logistic regression model using the predictions of the HDPMM topic model, the co-occurrence based model, the association norm based model, and a null model (where ϕ_i is set to 0 for all i), we calculate model fit statistics such as BIC, AIC, and Deviance. They are shown in the following table:

	HDPMM	Co-occur	Assoc	Null
BIC	5775.68	5824.33	6083.58	6212.77
AIC	5715.97	5764.62	6023.87	6186.23
Deviance	5697.97	5746.62	6005.87	6178.23

We will concentrate on the BIC results as the $\log_e$ of the Bayes Factor comparing any model $\mathcal{M}_0$ to model $\mathcal{M}_1$ can be approximated by half the difference of the BIC of models $\mathcal{M}_1$ and $\mathcal{M}_0$. Thus, the $\log_e$ of the Bayes factor comparing the HDPMM predictions to those of the co-occurrence based association model is 24.32. By any standard, this is overwhelming evidence in favour of the predictions of the HDPMM relative to those of the co-occurrence model. For example, Kass and Raftery (1995) argue that a log Bayes factor on a $\log_{10}$ scale that is greater than 2.0 is already *decisive* evidence in favour of the better model. In our case, our $\log_e$ result of 24.32 is 10.42 on a $\log_{10}$ scale. As the BIC of the association norm model is even greater than that of the co-occurrence model, there is overwhelming evidence in favour of the HDPMM relative to the comparison models.

For the recall memory task results, each set of recalled words by a participant on any given test j, which we will denote by $\omega_j = \omega_{j1}, \omega_{j2} \ldots \omega_{jn}$, can be reasonably viewed as draws from a subjective probability distribution that is the participant's memory representation of the contents of the text. We can calculate the likelihood of this data according to the probability distribution defined by any of our models, denoted generically by ψ, as follows:

$$\mathrm{P}(\omega_j|\psi) = \prod_{i=1}^{n}\prod_{v=1}^{V} \psi_v^{\mathbb{I}(r_i=v)} = \prod_{v=1}^{V} \psi_v^{r_{jv}}$$

where $\mathbb{I}(\cdot)$ is an indicator variable that takes the value of 1 if its argument is true, and r_{jv} is the number of times that word w_v occurs in ω_j, which in this case will be either $r_{jv} = 1$ if

[10] This is open-source software and is available at `https://github.com/lawsofthought/wilhelmproject`

[11] All raw data, and code for all analyses, can be found at `https://github.com/lawsofthought/gallium`.

word w_v was recalled and $r_{jv} = 0$ otherwise. The $\log_e$ of the likelihood of all the recall memory task data is

$$\log_e \prod_{j=1}^{L} \mathrm{P}(\omega_j|\psi) = \log_e \prod_{j}^{L}\prod_{v=1}^{V} \psi_v^{r_{jv}} = \sum_{j}^{L}\sum_{v=1}^{V} r_{jv} \log_e \psi_v.$$

These results are presented in the following table:

	HDPMM	Co-occur	Assoc
logLik	-14109.02	-15100.94	-16039.98
Deviance	28218.03	30201.88	32079.96

Given that the deviance is equal to the BIC plus a constant term, the difference of the deviances is identical to the difference of the corresponding BIC's. Approximating the $\log_e$ of the Bayes factor by half this difference, we therefore calculate a $\log_{10}$ Bayes factor for the evidence for the HDPMM predictions relative to those of the nearest model, the co-occurrence based association model, as 430.79. On the basis of the interpretation described above, this is again overwhelming evidence in favour of the HDPMM.

Discussion

In this paper, we have proposed — and then tested using a high powered behavioural experiment — a Bayesian account of how we form memories for spoken and written language. This account models how we use our background knowledge to form memories as a process of Bayesian inference of the statistical patterns that are inherent in each text, followed by posterior predictive inference of the words that are typical of those inferred patterns. We have implemented this model specifically as a HDPMM and applied it to an approximately 80m word corpus of texts taken from the BNC. This allowed us to make predictions of the probability of remembering any given word in each text from a sample of texts taken from the BNC. We tested these predictions in a behavioural experiment with 216 participants. The results of the analysis from both the recognition and recall data provided overwhelming evidence in favour of the Bayesian model relative to non-trivial alternative models.

References

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

BNC Consortium. (2007). *The British National Corpus, version 3 (BNC XML Edition)*. Oxford University Computing Services: http://www.natcorp.ox.ac.uk/.

Bower, G., Black, J., & Turner, T. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*(2), 177-220.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive psychology*, *3*(2), 193–209.

De Deyne, S., & Storms, G. (2017). *Small world of words, www.smallworldofwords.org.*

Gallo, D. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. Psychology Press.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114.

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, *101*(3), 371-395.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Hemmer, P., & Steyvers, M. (2009a). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Hemmer, P., & Steyvers, M. (2009b). Integrating episodic and semantic information in memory for natural scenes. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1557–1562).

Hemmer, P., & Steyvers, M. (2009c). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, *16*(1), 80-87.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association*, *90*(430), 773-795.

Kintsch, W. (1988). The role of knowledge in discourse comprehension - A construction integration Model. *Psychological Review*, *95*(2), 163-182.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, *9*(1), 111–151.

Nelson, D., McEvoy, C., & Schreiber, T. (2004). The university of south florida word association, rhyme and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 408-420.

Rawson, K. A., & Kintsch, W. (2002). How does background information improve memory for text content? *Memory & cognition*, *30*(5), 768–778.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385-407.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Steyvers, M., & Griffiths, T. (2007). Probabilisitic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Psychology Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.