Supplemental Materials for: Temporal Proximity links Unrelated News Events in Memory

Mitchell G. Uitvlugt & M. Karl Healey

Michigan State University

Supplemental Materials for: Temporal Proximity links Unrelated News Events in Memory

Date Assignment in Experiment 1

To begin, we compiled a list of major events leading up to the 2016 United States presidential election (e.g., candidacy announcements, debates, notable scandals, etc.). Keywords were then created for each event, and all subject-generated headlines that contained those keywords were grouped together and given the same date. For example, one commonly recalled story was the announcement that Donald Trump had won the election. Different subjects wrote different headlines to describe this single story (e.g., "Trump wins election in startling upset," "Clinton loses election," "Hillary Gives Concession Speech"). Keywords for this story were "Trump," "wins," "victory," "Clinton," "loses," "concession," "president-elect," "presidency," "election," "results."

Headlines that contained these keywords were provisionally grouped together, and then each headline in the group was read to confirm that it appropriately fit the given categorization. A date was then attributed to the story event (November 8, 2016 for the election result announcement example).

Some headlines related to recurring themes of the campaign and could not easily be associated with a single date. For example, many headlines described Donald Trump's plan to build a wall along the United States-Mexico border (example keywords: "Trump," "wall," "Mexico," "build"). This idea was first proposed at Trump's candidacy announcement, was emphasized throughout his campaign, and was a heated topic of discussion during the presidential debates. For these sort of headlines, we sought to determine the date when the story first received national attention (i.e., the date when the majority of people would have been first familiarized with it).

For this we used Google Trends—a website that provides data on when and how often specific words or combinations of words are searched on Google, effectively providing time-series data on the amount of public interest for any given topic. If we search for "Trump wall" on Google Trends within an appropriate time frame (`https://g.co/trends/vzBlr`), we see the first sharp increase in interest occurs in late

February, specifically on February 26, 2016 (`https://g.co/trends/L1BUI`). We then searched news articles from the day of the peak to corroborate that public interest did in fact peak here. In our example, the peak corresponded to a heated statement Trump made regarding the border wall at the Republican presidential debate in Houston.

Headlines that did not match any keywords were read individually. If the headline fell into a category but was simply missed by the keyword search, it was added to the appropriate group (e.g., "Trump Gets the White House" did not contain both "Trump" and "wins" but was still classified as a member of the "election result announcement" category and therefore was given the date of November 8, 2016). More obscure headlines were given dates on an individual basis (e.g., "Cher says she'll move to Jupiter if Trump wins" was given the date of June 16, 2015; when she tweeted her escape plan (`https://twitter.com/cher/status/610956742545911809`). Headlines that were too vague (e.g., poll results: "Hillary is pulling ahead"), an opinion (e.g., "Donald Trump will help all Americans"), or were altogether unrelated to the election (e.g., "Cubs win World Series") were removed from analysis. After eliminating these headlines, a total of 7,931 remained; on average 7.55 ($SD = 4.82$) per subject.

As a final step, we converted the calendar dates into day numbers to allow for easy calculation of the lag in days between pairs of headlines. Across subjects, the headline with the earliest calendar date was defined as day 1 and all other headlines were given values based on the number of days that separated them from day 1. To illustrate, the earliest headline submitted, "HRC has conference about emails and claims convenience and one device," occurred on March 10, 2015 (`http://cnn.it/198OsUn`) and was defined as day 1. The next earliest headline, "Hillary Clinton announces she's running for President," occurred 33 days later on April 12, 2015 (`http://politi.co/1akOaJN`) and was therefore assigned the day value of 34. The last headline recalled, "Trump won't pursue case against Clinton," referenced a statement Trump made on November 22, 2016 in which he reversed his campaign pledge to seek a new criminal investigation into Hillary Clinton (`http://politi.co/2fYU2MY`). This occurred 623 days after the first headline, and therefore it received a day value of 624.

Date Verification

To test the validity of our assignment of dates to headlines, we compared the subjects' responses from the ordering task with our date assignments. In the ordering task, subjects rank ordered all of their submitted headlines from the earliest to the most recent. To make our day values commensurate with these subject-provided rankings, for each subject we took the day values of their headlines and rank ordered them based on those values. For example, if a subject recalled three headlines with the day values 22, 1, and 333, they would be rank ordered as 2, 1, and 3.

As described above, we did not assign dates for some headlines because they were undateable. Subjects, however, had to rank all of their headlines, including those we could not date. If the subject indicated that a undateable headline occurred earlier than other dated headlines, then this could artificially increase the rankings of the more recent headlines. To fix this, we removed any such undated headlines and adjusted the subject-provided rankings as if they never had the opportunity to rank the undated headline. For example, imagine a subject recalled four headlines and rank ordered them in the following way: day 99, undated, day 32, and day 481. Initially, their subject-provided rank orders for those headlines would be 1, 2, 3, and 4, respectively. However, since their second headline could not be assigned a date, it would be removed from analysis, and our rank order would assign the following ranks: 2, – , 1, 3. To appropriately match our rankings with the subject-provided ranking, we removed the undated headline and adjusted the subject ranking to: 1, – , 2, 3.

If a subject recalled two headlines from the same day, we assigned them the same rank ordering. However, when subjects ordered their own headlines, they did not have the option to indicate that two events occurred on the same date. Therefore, in some cases, subject-provided rankings could be higher than our rankings. For example, imagine a subject ordered headlines from the following days: 22, 578, 611, 611. We would have rank ordered them as 1, 2, 3, 3, but the subject, who was not allowed to assign ties, would have had to order them as 1, 2, 3, 4. Therefore, a subject's highest rank could be higher than our highest rank.

If our date assignments are valid, then our rank ordering should be similar to the subject-provided rank ordering. Figure S1 shows there is indeed a strong positive correlation between our rank orderings and the subject-provided rank orderings, $r(7,577) = .667$, $p < .001$.

Semantic Similarity Ratings from Experiment 1

Events experienced nearby in time might be more similar than events experienced further apart. To control for this potential confound, we needed to measure the similarity between pairs of headlines. To accomplish this with the thousands of headline pairs, we utilized Amazon's Mechanical Turk to recruit independent raters to judge the similarity between headline pairs. Each rater viewed and scored a total of 40 pairs of headlines and was paid $1.00.

Each pairing consisted of two headlines that were adjacently recalled by a subject in the experiment. All transitions were rated, even if we were unable to assign a date to one of the headlines in the pair. The 40 pairs of headlines a rater saw did not come from the same subject; instead, they were randomly selected from all pairs available across all subjects. Each pair of headlines was presented to multiple raters but was never presented to the same rater more than once. Similarity was rated on a scale from 1 (not similar) to 10 (very similar). There was also a "Not Applicable" option below the numerical choices, which was to be selected if one (or both) of the headlines being rated did not relate to the 2016 election. If two or more raters scored a pair of headlines as "Not Applicable" the pair was given no similarity rating and that headline transition was removed from the semantic similarity analysis.

Before rating actual headlines, raters were presented with a series of example headline pairs to demonstrate appropriate similarity ratings. For example, we presented raters with two possible headlines: "Hillary Clinton Nearly Faints at 9/11 Memorial Site" and "Melania Trump copies Michelle Obama's speech at RNC." We then instructed raters that because these headlines described different events and focused on different people that this pairing should be assigned the lowest similarity rating. Raters

were also given an example of a headline pairing that should receive a high similarity rating: “Clinton loses” and “Trump wins.”

Rater Quality Control

To ensure high quality similarity ratings, we excluded raters based on the following three criteria. First, if a rater finished scoring all 40 headlines in 1 minute or less, all of their ratings were discarded. Second, we excluded raters who provided scores that were consistently different than other raters’ scores for the same headline transitions. To quantify this, for each headline transition, we averaged the similarity ratings from all the various raters who scored it. Then, for each similarity rating, we calculated how far it was from the group average for that headline. We averaged this value for all 40 transitions that a given rater evaluated, creating a score reflecting that specific rater’s deviation from the group average. Doing this for each rater provided a distribution of deviation scores. If an individual rater’s deviation score was more than 2.5 standard deviations away from the mean of this distribution, all of their ratings were discarded. Lastly, for each rater, we counted how many times they were the sole rater to score a given pair of headlines as “Not Applicable.” If a rater’s count was more than 2.5 standard deviations away from the across-rater mean, all of their ratings were discarded. After excluding raters based on these criteria, we checked whether each pair of headlines was still rated by at least four raters. If not, we had new raters score those remaining headlines, and then ensured that their ratings qualified for inclusion.

Temporal Bias Score

To visualize how much larger the observed contiguity effect was than would be expected by chance (i.e., the random-transition model), we calculated *temporal bias scores* for each subject.

Because lags so strongly clustered around lag zero (Figure 1), we began by binning lags into the following unequally spaced bins to better visualize the critical near-lag transitions: < -100, -100 – -51, -50 – -11, -10 – -1, 0, 1 – 10, 11 – 50, 51 – 100, and > 100. Then, for each bin, we counted the *actual* number of times each

subject made a transition falling into that bin and the *expected* number of times such a transition would be made under the random-transition model. To avoid any division by zero, the machine precision epsilon (i.e., the smallest representable number such that $1 + \epsilon \neq 1$) was added to the actual and expected counts of each bin. For each subject and each lag bin, we used these counts to calculate a temporal bias score which reflects the bias toward making transitions of that lag:

$$Temporal\ bias\ score = \frac{actual\ count - expected\ count}{expected\ count}. \quad (1)$$

Figure 3 in the main text shows the across-subject average temporal bias scores as a function of lag.

Details on the Latent Semantic Analysis procedure from Experiment 2

For each news event subjects recalled, they provided a url to story covering the event that was published as near as possible to when they first learned of the event. Form each of these urls, we extracted the text of the story. Each story was treated as a document in a corpus.

We subjected this corpus to a standard LSA pre-processing pipeline (for a turorial see, Landauer, Foltz, & Laham, 1998) using the Python packages Natural Language Toolkit (Bird, Klein, & Loper, 2009) and gensim (Řehůřek & Sojka, 2010), which are widely used in the Natural Language Processing literature. Specifically, we first removed extremely common words using the list of "stop" words from the Natural Language Toolkit (e.g., a, the, he, her, etc.), lematized the words (e.g., walked, walking, walker all become walk), removed any words that occurred only once in the corpus, and finally applied the term frequency–inverse document frequency (TFID) transform (Robertson & Jones, 1976) which estimates the importance of a given word to a given document by accounting for the fact that some words appear many times in particular documents (e.g., 'Trump" in political stories) whereas other words appear many time in many different kinds of documents (e.g., "email" in both political and technology stories).

Next, we subjected the TFID model to the standard LSA algorithm as implemented in the gensim package (Řehůřek & Sojka, 2010), which preforms a

Singular Value Decomposition. When creating the vector representations we retained 300 dimensions. Then for each pair of news stories, we computed the cosine of the angle between their 300-dimensional vector representations yielding a similarity matrix that is directly analogous to the similarity matrices used in free recall list learning studies.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (`http://is.muni.cz/publication/884893/en`)

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, *27*(3), 129–146.

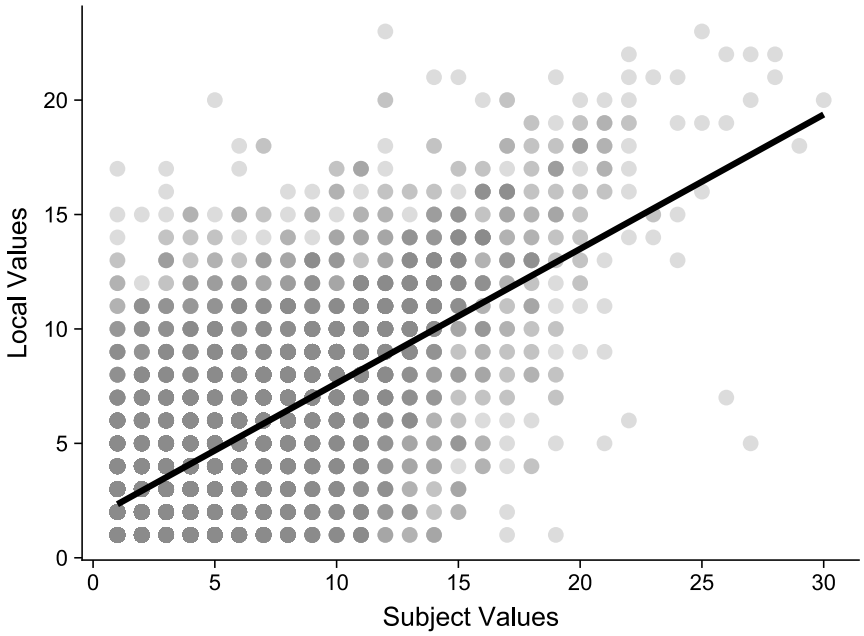

Figure S1. Relationship between subject values and local values. Each subject ordered their headlines in chronological order, creating our *subject value* variable. Similarly, within each subject, we rank ordered headlines based on our assigned day values, creating a *local value* variable. The figure shows the relationship between these two variables. The two sets of values are highly correlated ($r(7577) = .667$, $p < .001$).

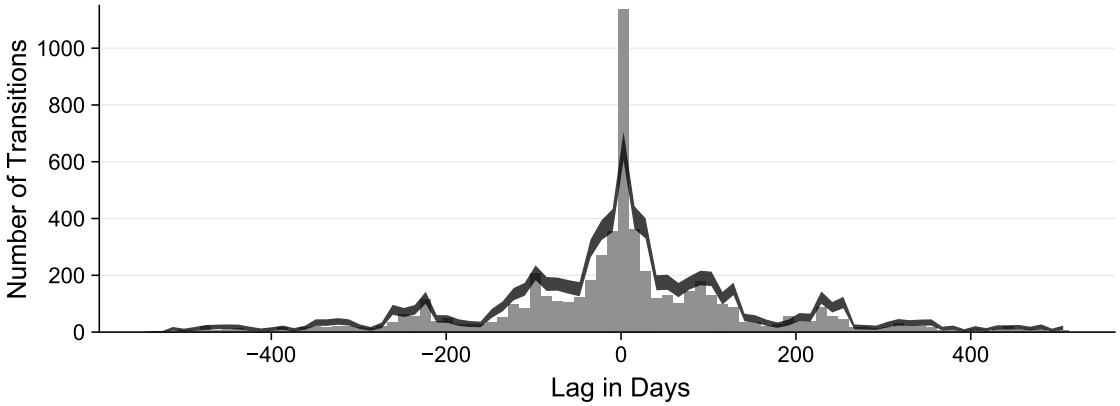

Figure S2. Distribution of transition lags in Experiment 1 with no subject exclusions. Subjects recalled as many news stories related to the 2016 United States presidential election campaign as possible, in whatever order they came to mind. Each time a subject recalled one story and transitioned to recalling another story, we defined the lag of the transition as the difference, in days, between when the two stories had originally appeared in the news. The light gray bars are a histogram showing that the distribution of these lags across subjects peaks at zero—subjects preferred to transition between stories that occurred near together in time. The darker gray line shows the middle 95% of the distribution of lags from a null model in which transitions are random; the null model has a lower peak than the actual data.