Incidental Verbatim Memory for Language

Olga Gurevich
Powerset, Inc.

Adele Goldberg
Princeton University
adele@princeton.edu
(609) 258 8772 (phone)
(609) 258 1113 (fax)

Submitted July 2008

1 Address correspondence to olya@powerset.com or adele@princeton.edu. The work described here was done while Gurevich was a postdoctoral fellow at the Educational Testing Service in Princeton, NJ and at Princeton University. We are grateful to NSF grant #0613227 to Goldberg for sponsoring this research. We are also very grateful to Charlotte Wieskittel for her senior thesis work on this topic (Weiskittel 2006), and to Devin Casenhiser for helping us run subjects. Charlotte ran the first recall study, using the Spiderman story, and began the collection of data on Gramps as well. Finally, we are grateful to Greg Murphy and Morten Christiansen for helpful feedback on an earlier version of this paper.
Abstract

It is widely believed that verbatim memory for language is virtually nonexistent except in certain circumstances, for example if participants are warned they are to receive a memory test, if the language is “interactive” (emotion-laden) in some way, or if the texts are exceedingly short and memory is tested immediately. At the same time, there has been a great deal of work in linguistics and psycholinguistics that has demonstrated that we retain a great deal of item-specific information about words, idioms and phrasal constructions. The present experiments revisit the question of verbatim memory for language and demonstrate that participants do reliably recognize and recall full sentences that they are exposed to only once at above chance rates. The texts are 300 words long, non-interactive, and no advanced warning of a memory test is given. Verbatim memory is demonstrated even when lexical content and memory for gist are controlled for, and even after a 2-day delay.

Keywords: verbatim memory, recognition, recall, usage-based model
Speakers are at once impressively creative and impressively repetitive. The creative aspect of language allows speakers to express new ideas in new contexts; this aspect of language has been emphasized within much of theoretical linguistics for over 50 years.

Other work has emphasized the extent to which speech consists of formulaic or prefabricated phrases (“prefabs”) (Pawley & Snyder, 1983; 2002). The formulaic aspect presumably simplifies the speakers’ task, in that many utterances or parts of utterances can be pulled directly off the mental shelf without having to build each utterance wholly anew, from scratch. In one study of a two year old child’s spontaneous speech, 63% of the child’s utterances were found to be verbatim repetitions of the child’s own earlier utterances (Lieven, Behrens, Speakers, & Tomasello, 2003). It turns out that children are also faster and more accurate at repeating utterances that occur with high frequency when lexical frequency and length are controlled for (Bannard & Matthews, 2008); the same is true for adults (Bod, 1998; Reali & Christiansen, 2007a, 2007b). This would seem to require that something about high frequency utterances or the way they have been put together is retained in memory in some form.

Oftentimes utterances are at once novel and formulaic in that they involve constructions that dictate much of the lexical content, while filling open slots of the construction with new lexical content. For example, even in young children’s speech, nouns are freely substituted for one another from very early on (Tomasello, Aktar, Dodson, & Rekau, 1997). Open slots of various types exist in all phrasal constructions that are not completely fixed idioms. Possessive phrases are often not fixed in advanced: one can jog his memory/her memory/their memory/the dog’s memory. Verb tense is also often flexible: one can commit a phone number to memory or have committed it to memory. One can take a stroll or a trip down memory lane (but cannot visit memory lane). These idioms can be represented as in (1)-(3), where the verbal lexeme is represented by capital letters to indicate that tense is not specified.

(1) JOG <someone’s> memory
(2) COMMIT <something> to memory
(3) TAKE a {walk, stroll, trip} down memory lane.

Clearly both aspects of language, its creative and its formulaic character, are important to understanding our great facility with language. Largely due to a recognition of this, there has been a growing convergence on “usage-based” models of language. These models are based on the idea that knowledge of language consists of a network of form-function correspondences at varying levels of specificity. Particular languages are learned by generalizing over utterances that a learner has heard used, while language production and comprehension involve combining or decomposing an utterance into its more basic form-function correspondences (Abbot-Smith, Dittmar, & Tomasello, 2007; Alishahi & Stevenson, to appear; Barlow & Kemmer, 2000; Bybee, 1985; 2004; Goldberg, 1999, 2006; Langacker, 1988; Lieven et al., 2003; Tomasello, 2003; Verhagen, 2002).

Just how specific is speakers’ knowledge of language? Memory for any experience is necessarily partially abstract insofar as the experience is not recorded completely. We might remember seeing a kumquat but we have abstracted away from the color of the
kitchen table upon which it sat; we also may have not noticed the tiny scratch in its surface or the exact length of its stem. So our mental representation of an experience, no matter how vivid, is partially abstracted from the actual experience itself. The same must be true for any memory of language.

In addition, usage-based researchers do not claim that every phrase or utterance is retained in memory. Most researchers assume that phrases are only retained if they have been witnessed with high frequency. Goldberg (2006:5), for example, has stated that “patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.” Yet here a logical problem arises. It is not possible to record that a pattern has been heard more than once if it is not recorded that the pattern is heard once. That is, “sufficient frequency” cannot involve some high number, call it n, of instances unless speakers retain some memory trace of an instance that has occurred with a frequency of n-1 so that the frequency can be increased upon subsequent exposure so as to tally n; and yet a frequency of n-1 cannot be recorded unless a frequency of n-2 was recorded. Applying this reasoning recursively, we must conclude that speakers would need to record a single exposure (frequency of 1), implicitly if not explicitly, at least for long enough that a subsequent instance might be encountered so that the frequency count could be increased (see also Bybee, 2006). If absolutely no memory trace were retained upon a single exposure, then each exposure would be the same as the very first exposure, and frequency (relative or absolute) would never accrue. No constructions would ever have “sufficient frequency.”

Thus if high frequency utterances are actually retained, some trace of some individual utterances must be retained for some period of time. Memory traces that are reinforced within a certain temporal window could become further entrenched and so would be less prone to quick decay (cf. Savage, Lieven, Theakston, & Tomasello, 2006 and for discussion of this idea in the non-linguistic domain, see Rovee-Collier 1995).

The idea that verbatim utterances need to be retained for a non-trivial time window highlights an important apparent paradox, which we might label the “Learning from the Forgotten” paradox. Conventional wisdom holds that, except in limited circumstances, people almost immediately forget the surface details of the language they hear and only remember the “gist.” This idea is an old one (Bartlett, 1932; Binet & Henri, 1894) and has received support from several studies that seemed to demonstrate that verbatim memory is lost as soon as an utterance has been understood (Anderson, 1974; Gernsbacher, 1985; Johnson-Laird & Stevenson, 1970; Sachs, 1967).

In a classic study by Sachs (1967), subjects heard a series of passages and were tested at varying intervals for their level of retention of sentences embedded within the passage. Sachs reported that “the original form of the sentence is stored only for the short time necessary for comprehension to occur.” She thus concludes, “the meaning of the sentence is derived from the original string of words by an active, interpretive process. The original sentence which is perceived is rapidly forgotten, and the memory then is for the information contained in the sentence” (Sachs, 1967, p. 442).

In another experiment, Bransford and Franks (1971) asked participants to recognize which sentences they had heard before by choosing among a new set of closely related sentences. All of the sentences presented during the experiment were combinations of component parts of a series of complex sentences or ideas. They found
that participants “spontaneously integrate the information expressed by a number of non-consecutively experienced (but semantically related) sentences into wholistic, semantic ideas” and thus participants falsely “recognized” novel sentences that contained the combined meaning of multiple individual sentences (1971, p. 331).

Bock and Brewer (1974:841) argued on the basis of subjects’ tendency to falsely recall related but unwitnessed sentences that “an abstract representation of the meaning was remembered rather than the exact words, and…in recall the surface structure was reconstructed from this abstract representation” (emphasis added). This idea is echoed in much early work, e.g., “Listeners do not ordinarily retain the syntax of a sentence for longer than is necessary to grasp its meaning” (Johnson-Laird, Robins, & Velicogna, 1974). This idea remains the current wisdom on the topic: e.g., “Research on memory for verbal materials has demonstrated that sentences are quickly transformed into an underlying abstract meaning and that the original surface structure is lost” (Holtgraves, 2008:361).

**Surface syntax is remembered in certain circumstances**

It is recognized that surface syntax can be remembered under certain limited circumstances, in particular: when subjects are warned they will have to remember the sentences verbatim; when the utterance is emotion-laden; or when sentences are not presented as part of a coherent text.

Memory for surface form has been demonstrated to exist when subjects are told in advance that they will be tested on this information, particularly when the texts consist of only a few sentences. In a recognition task, Johnson-Laird and Stevenson (1970) systematically varied whether subjects anticipated a memory test would follow exposure to a short orally presented passage (exactly how short the passage was is not stated); they found that only those who were warned that the memory test would follow demonstrated above-chance performance. They conclude by noting that “The results suggest that subjects tend to retain syntax…only if they know that they are to receive a memory test (cf. also Johnson-Laird et al., 1974). The authors go on to suggest that some form of verbal rehearsal may underlie the above-chance scores when the memory test was anticipated. Such rehearsal would clearly be an atypical comprehension strategy. Reyna & Kiernan (1994) also found evidence of verbatim memory, in six and nine year old children; however, they too warned participants and tested them immediately after having heard each of 8 three-sentence texts. The immediacy of the testing, the brevity of the texts involved and the forewarning involved make these studies ill suited to addressing the question of whether subjects retain verbatim memory in more naturalistic contexts.

We have all had the experience of having a particular hurtful, flattering, colorful or funny comment “ring in our ears” weeks or months after the event. In fact, verbatim recognition memory has been demonstrated for such “highly interactive” utterances. Murphy & Shapiro (1994) had one group of subjects read a rather bland, non-emotional letter, while another group read a biting, sarcastic letter. They found that subjects were better able to remember a given sentence when it received the sarcastic interpretation
than when it received a bland interpretation in the context of the other letter. Keenan, MacWhinney, & Mayhew (1977) likewise tested the memory of subjects exposed to an academic lecture after a 30 hour delay. In a recognition task, the experimenters found that listeners were better at retaining surface information for utterances with a high “interactional” content in a recognition task than they were for utterances with a low interactional content. However, the Learning from the Forgotten paradox is not solved if only these sorts of utterances leave any sort of memory trace. Our linguistic experience is full of more mundane utterances, and on the usage-based view, we learn generalizations over these as well.

Finally, researchers have documented that verbatim memory is improved when subjects are exposed to isolated sentences or unrelated items that do not form a semantically coherent passage (Anderson & Bower, 1973; Peterson & McIntyre, 1973; Villiers, 1974). Gernsbacher (1985) suggests that integrating sentences into a coherent semantic representation seems to be causally involved in the loss of verbatim memory, insofar as coherent texts facilitate the creation of a summary gist or mental model. But again, this laboratory-induced phenomenon does not reflect our normal experience with language. While input is occasionally perceived as incoherent, we are quite adept at imbuing conversations with coherence, even when the utterances appear to be unrelated (Grice, 1975).

Thus none of the recognized ways in which verbatim memory is known to be enhanced are particularly helpful for solving the Learning from the Forgotten paradox. If people learn and use language on the basis of generalizing over stored utterances, then they must retain some memory trace of utterances long enough to hear another similar utterance, even when they are not warned they will receive a memory test, and when the utterances are heard in ongoing coherent discourse. In natural contexts, language users are not asked to remember sentences verbatim, and they are almost always exposed to discourse that is semantically coherent. Moreover, language users should retain some memory trace of at least some utterances that are not particularly emotion-laden, since users learn and use non-“interactive” utterances much of the time.

If we are to consider the theoretical position that formal patterns emerge from experience to be tenable, experience with language should leave some sort of memory trace, although memory traces need not be explicit (see below), and they may be partially abstract due to selective encoding. They can also be expected to decay if not reinforced over a non-trivial time window.

Kintsch and Bates (1977) represents the rare study that has demonstrated some verbatim recognition memory in a naturalistic context. Subjects listened to a regular lecture in which, unbeknownst to them, a subset of stimuli sentences were embedded. In the following class, two or even five days later, subjects were given a recognition memory test on the set of sentences and found to recognize old sentences significantly better than paraphrases. The old sentences and paraphrase sentences had overlapping but distinct open-class lexical items, however, leaving open the possibility that the

---

2 Whether or not subjects were significantly above chance at verbatim recognition in the “bland” condition is not stated, although they were said to correctly identify old sentences 71% of the time, while incorrectly labeling paraphrases as old 54% of the time, which is suggestive of some verbatim recognition.
appearance of verbatim memory of sentences was driven by memory for particular words; we return to this point when experiment #2 is introduced. The studies reported here collectively serve to support and extend this early finding in a number of ways.

*Implicit verbatim memory*

An argument can be made that the Learning from the Forgotten paradox is resolvable by appeal to *implicit* memory of language. Implicit memory differs from explicit memory in that it persists without conscious awareness, and in fact persists even in the face of general amnesia. Language clearly involves implicit memory (at least to a great extent) insofar as amnesics generally do not forget their native tongues even when they have forgotten their spouses and their hometowns.

Evidence of implicit memory comes from differential performance on tasks due to exposure, despite a lack of conscious awareness that memory has been accessed. For example, after reading a list of words including *church*, a subsequent implicit memory task could involve asking subjects to come up with any word that could be completed as follows: “c h _ _ _ _.” An explicit memory task would ask subjects to *remember* which word on the list could be completed as “c h _ _ _ _.” The word-stem completion task is the same, but the first instruction requires only implicit memory, while the latter instruction requires explicit memory. Amnesic patients perform the same as control subjects on the implicit word completion task, showing an increased likelihood of supplying words that they had been exposed to; but they are severely impaired relative to controls when the word completion task involves explicit instructions to recall words from the list (Graf, Squire, & Mandler, 1984).

It is widely recognized that we often retain implicit memory for language (and other stimuli) without having explicit awareness. For example, in one study, normal participants who took part in an auditory divided attention task heard related pairs involving low-frequency homophones such as taxi-FARE in an unattended auditory channel. Although participants were at chance on an explicit recognition task, they were more likely to use the low frequency spelling “fare” than “fair” when asked to spell a list of homophonous words subsequently (Eich 1984).

Beyond the word level, the phenomenon of structural priming indicates implicit memory for structural patterns. That is, exposure to instances of one grammatical pattern primes speakers to produce or comprehend other instances of the same. Kathryn Bock and colleagues have shown in a number of experimental studies that, for example, passives prime passives; ditransitives prime ditransitives, and datives prime datives (Bock, 1986; Bock & Griffin, 2000; Bock & Loebell, 1990; Bock, Loebell, & Morey, 1992; cf. also Branigan, Pickering, Liversedge, Stewart, & Urbach, 1995; Chang, Dell, Bock, & Griffin, 2000; Garrod & Pickering, 2007; Griffin & Weinstein-Tull, 2003; Hare & Goldberg, 1999; Kaschak, 2007; Loebell & Bock, 2003; Saffran & Martin, 1997; Savage, Lieven, Theakston, & Tomasello, 2003; Yamashita, Chang, & Hirose, 2003).

Structural priming has been demonstrated to last over 10 intervening filler trials (Bock & Griffin, 2000; Chang et al., 2000; Huttenlocher, Vasilyeva, & Shimpi, 2004), and after a delay of 20 minutes (Boyland and Anderson 1998). Implicit memory for individual sentences (and structurally related sentences) can span 40-50 intervening sentences and an additional intervening five-minute distracter task (Luka and Barsalou
At the same time, the sentence lists in these experiments do not involve coherent texts, a situation that could artificially enhance memory of utterances, as discussed above.

Nonetheless, implicit knowledge of language may clearly be helpful in addressing the Learning from the Forgotten paradox insofar as the requisite item-specific knowledge might not be forgotten even if it is not consciously accessible in recognition or recall tasks. Previously heard utterances could be implicitly retained and could serve as the basis for implicit generalizations. This could potentially make similar types of sentences easier to comprehend or produce without any conscious awareness of the types of sentences that had been heard before.

Yet if we want to account for speakers’ full ability to use their knowledge of language, we must recognize that important aspects of that knowledge must be accessible to consciousness. People generally have reasonably good metalinguistic awareness of particular words and idioms. For example, we know that we’ve heard utterances such as those in A often, while we’ve rarely heard utterances such as those in B in Table 1:

[INSERT TABLE 1 HERE]

We have conscious awareness of a vast number of such idioms and prefabricated utterances. We are able to consciously adjust our choices of lexical items, idioms and constructions so that they are used in appropriate semantic and social contexts. The type of semantic memory required is generally accepted to be a subtype of explicit memory: we need to be able to recall and recognize the meanings of words, idioms and prefabs. People are also able to explicitly judge the relative frequencies of comparable items to an impressive extent (Hasher & Zacks, 1984; Shapiro, 1969).

Moreover, while much long-term knowledge of language is clearly implicit, the initial learning of language may well require access to explicit memory. Anterograde amnesiacs do not learn new languages or even new words. The famous amnesiac H.M. lost his memory due to drastic brain surgery for epilepsy in 1953, and he was subsequently unable to learn the words Jacuzzi, hippie or granola\(^3\) (Corkin, 2002). Unimpaired people, on the other hand, are often able to explicitly recognize and use a new word after a single exposure (Carey & Bartlett, 1978). The ability to retain new experiences in long-term memory appears to be a prerequisite for second language learning (cf. also Ellis, 2002; Knowlton, Ramus, & Squire, 1992; Schmidt, 1990).

The relationship between implicit and explicit memory is complicated and there may well be various levels of explicit memory (Karmiloff-Smith, 1986) and different

\(^3\) That is, Corkin (2002) demonstrated that there is no facilitation in implicit word-completion tasks for this type of unfamiliar word in the case of the amnesic patient H.M. At the same time, (Glisky & Schacter, 1986) were able to teach amnesics new words by repeated exposure to the word involving slowly vanishing cues; this technique capitalized on subjects’ intact implicit memory but clearly represented an unusual word-learning situation, one that is not duplicated in naturalistic contexts.
types of implicit knowledge; moreover, implicit and explicit may be more closely related than is sometimes assumed (Turk-Browne, Yi, & Chun, 2006). In this paper we sidestep these important issues in that our aim is a simple one. We revisit the long-standing issue of whether explicit verbatim memory for language is above chance in naturalistic contexts of language use: when speakers are not warned that they will receive a memory test and when sentences are presented as part of coherent texts and are not interactive or emotion-laden. We investigate verbatim memory in both recognition (experiments #1-2) and recall of texts (experiments #3-5). Experiments #2 and #4, moreover, control for possible lexical memory effects. In a final study (#5), we investigate whether above chance recall exists after a 2-day delay.

Previous results that had been offered as evidence against explicit verbatim memory are reconsidered in light of the present results. We find that various factors have conspired to downplay the importance of verbatim memory in past research. When looked at in depth, previous work can be seen to support the view that verbatim memory for language exists. The existence of some verbatim memory, lasting for some non-trivial amount of time, provides a strong indication that a usage-based approach to language is indeed viable.

Criteria for addressing the Learning from the Forgotten paradox

Our experimental designs satisfy four important criteria. First, they involve realistic, connected text since lists of words or other disconnected text are not likely to be processed and remembered in the same way as natural discourse. We use text and illustrations from a children’s storybook as stimuli, which allow us to approximate natural speech, as both natural discourse and children’s books primarily use simple sentence patterns, shorter sentences and higher frequency words. One of the storybooks was written as a first person narrative and the other included dialogue, both of which are standard forms in spoken language.

A second criterion was that participants were given no indication that their memory of the text would be tested. Speakers’ naiveté was confirmed during debriefing. Forewarning was avoided because it could cause subjects to rely on a special purpose encoding strategy instead of allowing comprehension to occur naturally.

A third criterion was that the texts were long, at around 300 words. This ensured that artificial forms of encoding were unlikely; the texts were too long to support auditory or other forms of conscious rehearsal, for example. The longer texts also ensured that memory would be tested more than a few syllables downstream.

A final criterion was to have a way to determine, in the case of free recall (experiments #3-5), whether subjects were actually remembering texts verbatim or whether they simply happened to formulate the same content in the same way as the author of the story had. Insofar as language is formulaic, this is a real concern. That is, while it is possible to greet someone with any one of a number of possible utterances including, how are you?, hello, nice to see you, you look familiar, these phrases could be uttered due to priming by an interlocutor or they could just happen to be produced. To control for this issue, we created two versions of each story; each clause from version 1 had a corresponding clause in version 2 with the same content but different surface form. The research question was therefore whether subjects who heard a particular formulation were more likely to use that formulation than subjects who had heard a paraphrase of the
same content. This design allows us to distinguish verbatim memory from memory for gist.

The next two sections describe two experiments using different stories, designed to assess verbatim recognition of a coherent, fairly long and non-“interactive” text, for which no warning that memory would be tested was given.

**Experiment #1: Spiderman recognition**

Experiment #1 involves a visually presented recognition task of an orally presented naturalistic story. Subjects are presented on one of two versions of the story and are then tested on clauses that appeared in either of the two versions. Their task is simply to decide if the item had been witnessed before (old) or whether it is new (paraphrase). For each subject, half of the test items came from the version they had heard and the other half from the alternative version.

**Participants**

Twenty four Princeton University undergraduates participated in the study, fulfilling a mandatory experimental requirement for a psychology course.

**Materials**

The test materials were based on an illustrated children’s book, *Spider-Man: I am Spider-Man* (2002). The book is 32 pages long, 300 words, and each page contains its own picture and between one and four sentences of text. The text is written in the first person and does not include any dialogue.

While the pictures serve as cues for advancing the story, the accompanying sentences are not descriptions of the content of the pictures (see Appendix for the two versions of the story and the full set of pictures).

In order to test specifically for memory of surface syntactic structure, we created a second version of the story with the same content and the same word count. Sentences across the two versions were matched for the same gist content as those in the original story, but had different syntactic structures. Some examples of the sentences in the two versions are below.

*Story version 1*: My fingers can stick to anything.  
*Story version 2*: I can stick my fingers to anything.

*Story version 1*: I am strong enough to fight four bad guys at once!  
*Story version 2*: Fighting four guys at once is easy for someone as strong as me.

The pictures were scanned and the text was digitally removed. A male narrator recorded the stories. The total playing time was 4:02 minutes for the original version of the story and 4:12 minutes for the alternate version.

For purposes of subsequent analysis, we subdivided the two stories into clauses, with one main predicate per clause. The analysis only considers those clauses that differed by more than one word between the two versions. There were a total of 37 such clauses, although there were 54 clauses in each story, in total.
As a between-subjects variable, we varied whether pictures from the original text were used as prompts during the recognition task (cued-recognition: group 1a) or whether random pictures appeared before each prompt (non-cued-recognition: group 1b). We hypothesized that pictures from the text would prompt participants to reinstate the context, which might increase recognition accuracy. Therefore, in addition to the pictures from the comic book for group 1a, we used unrelated distracter pictures from the Internet, depicting various scenes of nature for group 1b.

Procedure
The experiment consisted of two stages: listening, and performing the recognition task. Before the start of the experiment, subjects were not warned about the purpose of the study. They were simply asked to listen to a story while looking at the accompanying pictures on a desktop computer.

In the listening stage, subjects heard one of two versions of the story while looking at the corresponding pictures. The story was advanced automatically, as soon as the accompanying sound was finished playing for each picture.

The recognition testing phase began immediately after the story was presented. Written clauses were displayed on the computer screen one at a time and subjects were asked to answer “yes” or “no” to whether or not they heard that exact clause in the story. Before each clause, a corresponding picture was shown for 1000 milliseconds; either the picture that had appeared when that content had been heard (in the case of group 1a subjects) or an unrelated picture (group 1b subjects). Subjects recorded their answers by pressing one of two keys. The selection of a key caused the program to progress to the next clause.

We created two mixed lists consisting of clauses from the two story versions, each containing half the clauses from one story and half from the other, chosen at random, thus assuring that half the clauses were from the story version they heard (i.e. matching, with a veridical answer of “yes”), and half were from the alternative version (non-matching, with a veridical answer of “no”). The clauses were presented in the same order as in the original story. Each subject was given one of the two lists in the recognition stage. In several instances, the same picture was accompanied by more than one recognition clause. In these instances, the corresponding distracter picture was also presented before each clause, such that the presentation was exactly parallel with either set of pictures.

Thus the design was mixed-factorial. The between-subject factors were 2 (story 1 or story 2) x 2 (original or distracter picture) x 2 (recognition clause list).

Results
The measured dependent variable was accuracy. In order to smooth the effects of some participants having inherently better memory than others, we aggregated our data by subject and used the by-subject data points in measures of statistical significance.

The mean percentage of correct responses with the original picture was 0.72, and the mean with the distracter picture was 0.73: a non-significant difference (one-way ANOVA, $F(1,22) = 0.22, p > 0.5$). Thus results in which pictures from the story were displayed and those with distracter pictures are collapsed below. The overall average correct rate for Story 1 was 0.72, and for Story 2, it was 0.73: this was also a non-
significant difference \((F(1,22) = 0.10, p > 0.5)\). Therefore the results for the two versions are also collapsed below.

An overall mean of 0.72 of all responses were correct (i.e. “yes” to a matching sentence and “no” to a non-matching sentence), with standard deviation of 0.08. If subjects did not distinguish verbatim matches from gist-matches, we would expect an at-chance correct response rate of 0.50. A two-tailed t-test comparing the overall correct rate with the chance rate shows a significant difference \((t(23) = 13.60, p < .001)\); thus subjects did show above-chance verbatim recognition memory.

Breaking down the data further, we can compare the probability of a “yes” response to a matching sentence (true positive) and a non-matching sentence (false positive) (Table 2):

[INSERT TABLE 2 HERE]

Using signal detection theory, we calculated the sensitivity rate, d-prime, for distinguishing matching and non-matching sentences. A d’ of zero would indicate that there is no difference between signal and noise in the subjects’ responses. In fact, the mean d’ across individual subjects was 1.42 (N=24). A one-sampled t-test demonstrated that this d’ is significantly greater than 0 \((t(23) = 14.08, p < 0.01)\).

Discussion

The results of experiment #1 reveal that verbatim memory in recognition was significantly above chance, even though the experimental context was not highly interactive, the text was a cohesive one, and subjects were not warned that they were doing a memory task. In debriefings, subjects confirmed that they did not know they would be asked to retell the story, and they had not attempted to memorize the story when they heard it.

One concern about Experiment #1 was that the effect could have been driven, at least in part, by lexical effects (Reyna and Kiernan, 1994). If a speaker who heard Story version #2 remembered that the word easy was used to describe a particular content, then he would be more likely to correctly recognize that the second and not the first utterance had occurred in the text:

**Story version 1**: I am strong enough to fight four bad guys at once!
**Story version 2**: Fighting four guys at once is easy for someone as strong as me.

This is a possible confound in the Kintsch and Bates (1977) experiment that had shown evidence for verbatim memory as well, as their old and paraphrase sentences varied in terms of open-class words as did ours in Experiment #1. In order to control for the effect of lexical items, Experiment #2 was performed. In the second experiment, two versions of a story differed only in closed class words (prepositions, determiners, copula verb, light verbs, auxiliaries and morphology); all open-class content roots in each story were identical. It is generally accepted that closed-class words are not significantly affected by additional exposure, most likely because they are so frequent that their base rate activation does not change measurably. The fact that the two story versions differed
only in closed class items should essentially eliminate lexical effects on the memory task. If we still find evidence of verbatim recognition memory, it is stronger evidence that memory traces for full utterances are retained.

Another potential issue with Experiment #1 stems from the fact that the Spiderman story was written in the first-person. This could be argued to increase the "interactivity" of the context. Also, although our subjects had not seen this particular comic book before, the general outline of the Spiderman story is well-known in popular culture, conceivably making it easier to remember and recognize verbatim sentences; alternatively, one could argue that because the Spiderman script was known it would be easier to create an overall gist without relying on verbatim recall.

The story used in experiment #2 was written in the 3rd person and the story line was unfamiliar to our subjects. The two versions used, one heard by each of two groups of subjects, differed only in closed class items. The use of a second story also guards against item-effects.

**Experiment #2: Gramps and the Fire Dragon recognition**

**Participants**
Twenty-four Princeton University undergraduates participated in the study, fulfilling a mandatory experimental requirement for a psychology course. None of the subjects in Experiment 2 had participated in Experiment 1.

**Materials**
We chose another children’s story of comparable length to Spiderman, with illustrations that would act as cues but that were not depictions of the text. The story was *Gramps and the Fire Dragon* (2000). Some of the two-page illustrations were reduced to just one page for size consistency, so there were total of 25 pages in the experiment. The story is told in third person but includes dialogue. As in Experiment 1, the words were digitally removed from the scanned pictures.

The original story consisted of 39 sentences and 272 words, which we initially divided into 40 clauses for coding purposes. As before, we also created an alternative version of the story, controlling for number of words and content. For example, the original text of story version #1 was, "But that only made the dragon angry.” In version #2, which we created, the sentence was, “But that only angered the dragon.” In other cases, we altered a sentence into a fragment (or vice versa), or changed conjoined verb phrases into two sentences. (See Appendix for the two versions of the story.)

The second version of the *Gramps* story had slightly more words than the original (N = 299 vs. N = 272). For purposes of analysis, we selected only those clauses that differed by more than one word in the two versions of the story, resulting in 37 clauses. However, the full story was recorded in each case so as not to interrupt the narrative flow.

Both versions of the story were recorded by a male speaker. The sound clips were presented with each picture using E-prime. Total running time for Story 1 was 3:02 minutes; for Story 2, it was 2:54 minutes.

We again presented one group of subjects with distracter pictures instead of the storybook pictures. The distracter pictures were the same as those used in Experiment 1.
Procedure
The method of presentation and testing was exactly the same as in Experiment 1.

Results
The measured variable was accuracy.

The mean correct response when subjects viewed the original pictures was 0.60; the mean correct response when subjects viewed the distracter pictures was 0.59. As in experiment 1, these are not significantly different \( F(1,886) = 0.74, p > 0.5 \), so the two groups are collapsed in the following discussion. The overall correct rate for Story 1 was 0.63, and for Story 2 it was 0.56. This was also not a significant difference \( F(1,22) = 0.17, p=0.69 \), and so these two are also collapsed below.

The combined mean for correct responses was 0.59. To make sure that the difference was not due to a few speakers’ unusually good memory, we performed the calculation aggregated by subject; This mean was significantly greater than the “chance” rate of 0.50, as shown by a one-sample t-test \( t(23) = 5.60, p < 0.01 \). We also compared the probability of a “yes” response to a matching sentence (true positive) and a non-matching sentence (false positive) (Table 3).

[INSERT TABLE 3 HERE]

The mean sensitivity measure \( d' \) was 0.70 across all subjects \( (N=24) \). A one-sample t-test demonstrated that it was significantly greater than zero \( t(23) = 5.25, p<0.01 \).

These results show that, as in experiment 1, subjects were consistently able to distinguish between sentences they heard verbatim and sentences that had the same content but different surface form.

In debriefing, subjects stated that they did not realize their memory would be tested, and they did not try to memorize the story as they were listening to it.

Discussion
In both experiments #1 and #2, recognition accuracy was above chance, and responses to verbatim clauses were responded to significantly faster than responses to paraphrases. Experiment 2 replicated Experiment 1 in demonstrating that there is verbatim recognition of texts that are fairly long (roughly 300 words) and involve non-interactive, coherent texts, even when subjects do not expect a memory test. In addition, experiment 2 demonstrates that the effect is not due simply to memory for individual open class items, as the same open class roots were used in both stories; this controls for a possible confound that exists in much previous work. Moreover, it is not necessary that subjects already have a general “script” for the storyline (as they may have in the case of the Spiderman story used in Experiment 1), nor is a 1st person narrative essential.

Verbatim memory in previous literature revisited
How are studies #1 and #2 to be reconciled with previous findings that argued for a lack of verbatim recognition memory? It turns out that in several of the early studies, there are in fact indications that verbatim structures are not completely forgotten, even though the papers tended to emphasize (and be remembered as having demonstrated) a lack of
verbatim recognition. It seems that on a time-scale of decades, there is some truth to the idea that only the gist (of the articles) is remembered. The present studies strengthen earlier findings in that our subjects received no forewarning; study #2 also controlled for possible lexical effects.

In the original study by Sachs (1967), although the findings were discussed by Sachs herself as providing evidence that verbatim memory was lost, there was actually some evidence for verbatim memory traces. At delays of 80 and 160 syllables, the difference between true and false positives (i.e. verbatim memory) was significant (reported as $p < .01$, and $p < .05$ respectively). At the same time, this previous evidence must be qualified in that subjects were instructed that they would be asked to recognize original sentences, and they were provided with four practice passages as a warm-up. As noted earlier, we know from the Johnson-Laird and Stevenson (1970) study that such instructions enhance performance. Also, subjects were not tested at delays longer than 160 syllables, thus providing no direct comparison with our 300-word stories.

Jarvella (1973) tested recall and recognition of sentences that came close to the end of 100-200 word segments of dialogue. The test sentences were followed by a sentence that either repeated the test sentence, implied it (as with a tag question), or was novel. Immediately after this last sentence, subjects were asked to recognize or recall the test sentence. Recognition was correct 96% of the time for repeated sentences, 82% for implied sentences, and 76% for novel sentences, with significant differences between the three. However, as in Sachs’ study, subjects were warned that they were participating in a memory task, and the time elapsed between exposure and test was quite short (the length of one sentence).

Zimny (1987; as described in Kintsch et al., 1990) constructed two alternative versions of several 150-200 word stories. Subjects were not explicitly warned that the task was memory-related, but should have been able to guess it since each subject was tested on multiple stories. After hearing a story, subjects were asked to recognize one sentence, which was presented verbatim, with surface changes (paraphrase), with meaning implied from the original, or with different meaning. Surface memory was measured as the difference in recognition between verbatim sentences and paraphrases. On immediate testing, verbatim recognition (hits) was at 74%, and paraphrase recognition (correct rejections) was at 50%, giving an average correct rate of 62% (similar to our findings for Gramps and the Fire Dragon). However, again, subjects knew memory would be expected. Zimny found that when subjects were tested after a 40-minute delay, the difference between true and false positives became insignificant, leading her to conclude that verbatim memory was lost. However, the foils included not only different surface form, but in many cases also different meanings; knowing from previous tests what kinds of foils to expect may have predisposed subjects towards meaning-based memory.

Johnson-Laird and Stevenson (1970) found no evidence of verbatim recognition memory unless subjects were warned they were to receive a memory test. But, as was more common at the time, the study involved few subjects (12) and only a single stimulus sentence.

More recently, Hartgraves (2008), quoted in the introduction as presupposing a lack of verbatim memory, actually provides some relevant data, although it is not analyzed explicitly for verbatim recognition or recall. Instead, the studies demonstrated that people
are quite sensitive to whether sentences perform speech acts; for example, there is an increased likelihood of falsely remembering a speech act verb after being exposed to a sentence that performs an (indirect) speech act than after a control sentence. In Hartgraves’ experiment #2, participants were asked to read 24 short scenarios, each one 3-7 sentences in length; without advanced warning, their memory for certain sentences was tested after a five minute delay. Although comparisons to chance are not provided, when participants were given a choice of four sentences including a verbatim sentence, a semantically matched foil, and two semantically distinct foils, and asked to choose an exact match, they correctly chose the verbatim sentence 78% of the time (experiment #2). As discussed more below, another group of participants also recalled 19% of the sentences heard verbatim (Hartgrave’s experiment #4), when cued only with a single word (naming the speech act involved); a discussion of recall memory follows in section 5.

To summarize, memory for content or “gist” has been emphasized in these studies, and in their later interpretations by others. Clearly, when compared directly, memory for meaning always trumps memory for structure, but there seems to be no solid evidence that verbatim memory disappears entirely, and in fact previous studies seem to offer numerical indications that some verbatim memory does remain. Experiments #1 and #2 demonstrate clearly that verbatim recognition is significantly above chance, even when subjects are not warned in advance that they will receive a memory test and when lexical effects are controlled for.

**Recall experiments**

Much less research has been done on the type of retention that occurs in recall, and yet recall is clearly required on the usage-based model, if we assume that (implicitly and often partially) recalled utterances are used in production. Since recall is generally thought to be more difficult than recognition, and since recognition of verbatim language has been believed to be poor or nonexistent, it has likely been taken for granted that recall memory for verbatim language does not exist, except possibly in the sorts of limited contexts outlined in section 1.1.

And yet it is important to consider the ability to recall verbatim language, because the absolute number of correct responses in recognition tasks is not terribly meaningful. The number of correct responses clearly can be made to vary depending on how close lure sentences are to target utterances. Lures that differ only in tiny inconsequential ways (such as using contractions or not) are likely to decrease the number of correct responses, while lures that are insufficiently similar to the target sentence could artificially inflate the number of correct responses.

One of the first researchers to test recall in natural discourse was Bartlett (1932), who had subjects read a story about an unfamiliar culture and then retell it. He did not quantify his results, but he found that subjects tended to change or forget details that were inconsistent with their cultural frame of reference. This research led to the formation of Schema Theory, which was intended to explain and account for the incompleteness of memory. The theory claimed that incoming information was reduced to an integrated set of only the most relevant and important information, and that all irrelevant information and details, including the specifics of structure, are lost.
Some verbatim recall has been documented in earlier studies where very specific cues were used, and where subjects were warned about the memory task ahead of time. Jarvella (1973) was able to prompt recall by accompanying test sentences by audio tones and asking for recall after one intervening sentence was heard. He found verbatim recall rates of between 26% and 51%, depending on the nature of the intervening sentence. However, there was no control for whether the subjects remembered the surface form of the sentence, or remembered the gist and just happened to express it using the same surface form as the original (see next section for a more detailed discussion of this issue).

In a more controlled study, Bock and Brewer (1974) constructed different surface forms for lists of 10 sentences. After hearing a list, subjects were asked to write down what they remembered given one-word cues for each sentence, which were neutral with respect to surface form. Subjects successfully recalled 41% of the sentences verbatim compared to 16% close paraphrases (the rest of the productions were errors or omissions). The difference is significant (p < 0.01), but the short, artificially constructed sentence lists and advance knowledge of the memory task make it difficult to infer the existence of general verbatim recall memory from this study.

As mentioned above, recent work by Hartgraves (2008) demonstrated that 19% of the sentences heard were recalled verbatim (his experiment #4). Participants were exposed to 24 scenarios, and then asked to recall a subset of those sentences when cued with the speech act that the sentence had performed as well as the speaker’s name. For example, for a target sentence such as I like your new coat, participants were cued with “Heather; compliment.” For the target What time is it? Participants were cued with “Jane; ask.” However, because only one version of each sentence was used and many of the target utterances were fairly formulaic, it is conceivable that subjects primarily remembered the gist and just coincidentally formulated the meaning in a way that matched the target sentences.

The studies described in the next three sections investigate verbatim recall by prompting participants to recall specific sentences of fairly long stories, without prior warning about the memory task. Two versions of each story were used, with each sentence of version #1 paired with a paraphrase of it in version #2; in this way we were able to compare matches for heard utterances with coincidental matches to the alternative version.

The pictures from the recognition experiments were used as cues. In addition to immediate recognition of experiments #3 and #4 below, subjects were also tested after a two-day delay, as described in experiment #5.

**Experiment #3: Spiderman recall**

This experiment is based on Charlotte Weiskittel’s senior thesis (Weiskittel 2006) done under the supervision of Goldberg.

**Participants**

Twenty Princeton undergraduates participated in the study, fulfilling an experimental requirement for a psychology course. All were native English speakers.
Materials
We used the same Spiderman comic book as in the recognition study. As before, two versions of the story were created in order to compare the number of verbatim matches with the number accidental matches to the alternative version of the story that the subject did not hear.

Method
Participants were tested one at a time. They were asked to listen to a story while looking at the pictures on the screen, without instructions as to what type of task would follow. Each participant heard one of the two story versions. Immediately after the entire story was finished, an instructional page was shown that asked the participant to “read the story back” as accurately as they could while looking through the original pictures. The precise instructions are in the Appendix.

The retellings were recorded and transcribed. We analyzed each clause produced by a participant as a Match (which could be an exact match or a 1 word off match), or a Non-Match. We feel our coding scheme was conservative, even though we included utterances in which a single word was changed in the final tally of verbatim Matches because the differences involved were very minor. For example, we included as Matches changes from I am to I’m, or from I can go places no one else can to I can go places that no one else can. We did not count as verbatim matches those productions that differed by two words, even if they were otherwise reasonably close in form to the target sentence. For example, the following productions were not counted as matches:

- Heard:
  - The kids at school were amazed at my new strength.

- Produced (counted as non-match):
  - The kids at school were amazed by my new-found strength.

- Heard in alternative story:
  - My new strength amazed my classmates.

- Heard:
  - I was just like everyone else.

- Produced (counted as non-match):
  - I used to be like everyone else.

- Heard in alternative story:
  - I wasn’t always a superhero.

In addition, we classified each clause uttered as to whether or not it constituted a Match (exact or one word off) to the alternative version of the story that had not been heard by the participant. This was done so that we could compare Matches to the heard story to accidental Matches to the alternative version of the story that had the same content.
Results

Results demonstrate clear evidence of verbatim recall. There were a total of 48 clauses in either version, of which we considered 43 (because the remaining five only differed by one word). Participants who heard story 1 produced on average 33.8 clauses, or 70% of those heard; those who heard story 2 produced 31.6 clauses, or 65.8%.

Overall, of the clauses produced, 22.3% (N=144) were verbatim matches to a heard clause. Since subjects produced fewer sentences than they had heard, this constitutes 16.7% of all heard clauses being reproduced verbatim. 6% of clauses produced were coincidental matches to the alternative (unheard) version of the story. For subjects who heard story version 1 (the original version), 29% of clauses produced were verbatim matches to heard clauses. This corresponds to 22% percent of all heard clauses being produced verbatim; by contrast, 11% of clauses from story 1 were coincidentally reproduced by subjects who did not hear it. Subjects who heard story 2 (the reformulated version) produced verbatim matches 15% of the time, recalling 11% of the story, and only 1% of clauses from story version 2 were coincidentally matched by subjects who heard story version 1.

These results are summarized in Table 4 and Figure 1 below.

We compared mean verbatim matches across speakers, looking for a difference between matches to the heard story vs. coincidental matches. An overall 2x2 ANOVA (match / coincidental-match x story version 1 / story version 2) revealed a significant main effect of match/coincidental-match ($F(1,16) = 12.14, p < 0.01$), demonstrating evidence of verbatim memory. There was also a significant effect of story version ($F(1,16) = 10.18, p < 0.01$), reflecting the fact the number of coincidental matches to story 1 was significantly greater than to story 2 ($F(1,18) = 54.98, p < 0.01$). There was no significant difference in correct matches between version 1 and version 2 ($F(1,18) = 3.22, p > 0.05$). There was also no significant interaction between match type and story version ($F(1,16) = 0.11, p > 0.05$), indicating that the amount of verbatim memory for the two versions was not statistically different.

We also considered whether verbatim memory was evident for each story version, taken independently, a setting a higher bar for the demonstration of verbatim recall. In fact, verbatim memory was evident in the case of both story versions. In particular, comparing the average number of correct matches to story 1 by the subjects who heard it (N=10) with the average number of coincidental matches to story 1 by the subjects who heard the alternative story (N=10), there was significant evidence of verbatim recall for version 1 ($F(1,18) = 7.40, p < 0.01$). Similarly, the number of correct matches to story 2 by the subjects who heard it (N=10) was significantly greater that the average number of coincidental matches to story 2 by subjects who did not hear it (N=10) ($F(1,18) = 8.40, p < 0.01$).
Discussion

Subjects showed a remarkable ability to recall verbatim language heard in a fairly long story (300 words) without receiving any warning or indication that their memories for the stories would be tested. In line with earlier studies of verbatim memory, the majority of what participants recall is not verbatim matches with what they had heard: they by no means remember the entire stories verbatim. But the fact remains that a significant amount of verbatim memory is retained.

The fact that there were more coincidental matches to story version 1 than to the second story version demonstrates that our use of two versions of the same story is an important aspect of the design. If an analysis were to only consider the number of verbatim matches without comparing verbatim matches to coincidental matches, it would be subject to the criticism that the matches may be simply the result of remembering gist and formulating it in a natural way. It is probably fair to say that the language of the published story is more natural than our amateur attempt to write a children’s story; the greater number of coincidental matches to the published version is humbling evidence of this fact. Yet importantly, both the subjects who heard story version 1 and those who heard story version 2 were more likely to produce verbatim matches to the story they heard than could be expected by the rate of coincidental matches to the same story version. This is strong evidence that subjects retain significant verbatim memory for the purposes of recall.

Experiment #4: Gramps and the Fire Dragon recall

As discussed in preface to experiment 2, it is useful to test memory for language on more than a single story. Moreover, the two versions of the Spiderman story involved some differing open class words and so it’s conceivable that recall was encouraged by memory for particular open class words. For these reasons, we attempted to replicate the recall findings of experiment 4 with the Gramps and the Fire Dragon story (used in experiment 2). Recall that in the Gramps story, all open class roots were shared across the two versions of the story.

Participants

20 Princeton undergraduate students participated for course credit. None of the subjects participated in any of the other verbatim studies.

Materials

One of the same two versions of the children’s book Gramps and the Fire Dragon used in the recognition study 2 were read to each participant.

Method

The setup was the same as with the Spiderman recall experiment. Students listened to a version of the story while looking at the pictures. At the end, they were asked to repeat the story as best as they could while looking at the original pictures in the same order. Their responses were recorded, transcribed, and coded for whether they were Matches or Non-Matches, as described above.


**Results**

We again found evidence of verbatim recall. There were a total of 39 clauses, of which we considered 37 (because the other two differed only by a single word). Participants produced on average 31 clauses, or 82% of the number heard. Overall, of the clauses produced, 14% (N=85) were verbatim matches to a heard clause. Since subjects produced fewer clauses than they had heard, this constitutes 11% of heard clauses being reproduced verbatim; by comparison, subjects produced only 4% coincidental matches to the alternative story they had not heard.

For story version 1, 17% (N=52) of clauses produced were verbatim matches. This corresponds to 14% of the clauses heard being reproduced verbatim, while subjects who had heard the alternative story version produced only 5% (N=20) coincidental matches to story version 1. For story version 2, 9% (N=33) were verbatim matches to the heard story; by comparison, subjects only produced 2% (N=7) incidental matches to story version 2 if they had heard story 1.

The results are summarized in Table 5 and Figure 2 below.

[INSERT TABLE 5 HERE]

[INSERT FIGURE 2 HERE]

We again compared mean verbatim matches across speakers, looking for a difference between matches to the heard story vs. coincidental matches. An overall 2x2 ANOVA (match / coincidental-match x story version 1 / story version 2) revealed a significant main effect of match/coincidental-match (F(1,16) = 20.48, p < 0.01), replicating the finding of verbatim memory in experiment #1. There was also a small but significant effect of story version (F(1,16) = 6.24, p < 0.02), again reflecting the fact that the number of coincidental matches to story 1 was marginally greater than to story 2 (F(1,18) = 4.21, p = 0.06). The difference in correct matches between version 1 and version 2 was not significantly different (F(1,18) = 2.91, p > 0.05). There was also no significant interaction between match and story version (F(1,16) = 0.22, p > 0.05), indicating that the amount of verbatim memory for the two story versions was not significantly different.

We also considered whether verbatim memory was evident for each story version, taken independently, and we found again found that the effect was evident for each story version. That is, comparing the average number of correct matches to story 1 by the subjects who heard it (N=10) with the average number of coincidental matches to story 1 by the subjects who heard the alternative story (N=10), there was significant evidence of verbatim recall for version 1 (F(1,18) = 9.25, p < 0.01). Similarly, the number of correct matches to story 2 by the subjects who heard it (N=10) was significantly greater that the average number of coincidental matches to story 2 by subjects who did not hear it (N=10) (F(1,18) = 12.62, p < 0.01).
Discussion

Experiment 4 replicated the main finding of Experiment 3, namely that subjects’ tendency to repeat utterances from the story they heard when asked to retell it was more successful than could be expected by coincidence. Subjects demonstrated a significant tendency to remember parts of a story they heard once verbatim; again, as in Experiment 3, the story was fairly long and no warning of a memory test was given. Experiment 4 moreover ruled out the possibility that the verbatim recall in Experiment 3 was due to memory for open class lexical items, since open class items were shared between the two versions of the story used in Experiment 4.

Experiment #5: Spiderman recall after a delay

The previous experiments demonstrated evidence of verbatim memory in naturalistic recall conditions, which is consistent with the findings in Experiments 1 and 2 involving recognition paradigms. The recall studies involved a delay over the time it takes to tell the story, roughly 3-4 minutes. This interval is considerably longer than what would be considered necessary for comprehension to occur, but not so long as to demonstrate long-term memory effects. In order to address the Learning from the Forgotten paradox, it is necessary that verbatim memory should last for more than a few minutes. That is, if participants are able to record any information about the frequency of use of such utterances, a second repetition that would strengthen the memory trace cannot be expected to occur immediately.

In order to find out more about how recall memory may decay, we conducted a similar recall study where participants were asked to recall the Spiderman story after a 2-day delay.

Participants

Twenty Princeton undergraduates participated in a two session the delayed recall study. They participated in the first session for course credit, and participation in the second session was voluntary either for more course credit or for a $10 fee. Most students opted for the extra course credit.

Materials

The same materials were used as in the immediate recall study with the Spiderman story (Experiment # 3).

Method

The first session was identical to the immediate recall study. After completion of the recall task, students were asked to come back two days later to answer some questions about the stories they heard that day, without specifying which particular story and again without explicit warning about a memory task.

In the follow-up session, participants were again asked to recall the story while they were shown the original Spiderman pictures. They were not given time to look at the pictures first, and did not hear a replay of the recorded story. The amount of delay depended on a student’s schedule and varied between 1 and 3 days after the initial session, with the majority of participants coming back after 2 days.
Results

We analyzed the results from both the immediate and the delayed recall tasks. The immediate recall task replicated the initial recall experiment. Results for immediate and delayed recall are summarized in Table 6 below.

[INSERT TABLE 6 HERE]

In the immediate recall task, overall, 26% (N=151) of clauses produced were verbatim matches to the original story version, corresponding to 18% of clauses being reproduced verbatim, while 3% were coincidental matches to the alternative version of the story. Breaking those numbers down by story version we find that for story version 1 (the original book version), 31% of clauses produced were verbatim matches, or 21% of clauses reproduced verbatim, while 5% were coincidental matches. For story version 2 (the reformulated version), 21% of produced clauses were verbatim matches, or 14% of clauses were reproduced verbatim, while 1% were coincidental matches.

Replicating results from experiment #3, a 2x2 ANOVA (match/non-match x story version 1/story version 2) revealed a significant main effect of match / non-match ($F(1,16) = 24.63, p < 0.01$), a small but significant effect of story version ($F(1,16) = 4.89, p = 0.034$), and no significant effect of the interaction between the two factors ($F(1,16) = 0.14, p > 0.05$). As in our original recall study, there was no significant difference in number of matches for the two story versions ($F(1,18) = 1.78, p > 0.1$). However, the number of coincidental matches to story 1 was significantly greater than for story 2 ($F(1,18) = 15.23, p < 0.01$).

The number of verbatim matches to story version 1 produced by the subjects who heard it (N=10) was significantly greater than the number of coincidental matches to it produced by the subjects who heard story version 2 (N=10), $F(1,18) = 20.83, p < 0.01)$. Similarly, the number of verbatim matches to story version 2 from subjects who heard it was significantly greater than the number of coincidental matches to it by subjects who did not hear it ($F(1,18) = 8.00, p < 0.05$).

In the delayed recall task, overall a total of 17% (N=108) of the produced clauses were verbatim matches to the original, corresponding to 13% of heard clauses being reproduced verbatim, while 3% were matches to the alternative story.

An overall 2x2 ANOVA of the delayed recall task (match / non-match x story 1 / story 2) revealed a significant main effect of match/non-match ($F(1,16) = 20.79, p < 0.01$), demonstrating verbatim memory after a 7-day delay. There was no significant effect of story version ($F(1,16) = 3.23, p > 0.05$) and no significant interaction between match and story version ($F(1,16) = 0.37, p > 0.05$). There was no difference in accuracy between the two story versions ($F(1,18) = 1.55, p > 0.1$) and a small but significant difference in the number of coincidental matches ($F(1,18) = 6.15, p < 0.025$).

Verbatim recall was significantly greater than the coincidental matches for each story version taken independently, as well: for story version 1, $F(1,18) = 14.19, p < 0.01$; for story version 2, $F(1,18) = 7.37, p < 0.02$, and overall, $F(1,38) = 19.92, p < 0.01$.

As a repeated-measures statistic by subject, the difference between immediate and delayed recall was significant ($F(1,38) = 12.15, p < 0.01$), suggesting that verbatim memory decayed over time, but not to the point of disappearing completely. The
difference between the number of coincidental matches to the alternative story was not significant between immediate and delayed recall ($F(1,38) = 1.05, p > 0.5$).

Since participants were asked to recall the story immediately after hearing it as well as after a delay, it is possible that the delayed memory is conditioned, or at least facilitated, by memory of the initial recall. In other words, participants may be remembering what they themselves said rather than (or in addition to) what they heard.

To try and tease apart these two hypotheses, we coded the delayed recall responses as to whether or not they corresponded to the initial responses given by the same participant, regardless of whether they were Matches to the heard story version. 27% were Matches to what they had said before ($N=178$). But subjects were much more likely to repeat sentences that they had said before when those sentences had actually occurred in the original story than when they had not. In particular, 52% ($N=79$) of Matches produced in the immediate test were reproduced as Matches after the delay, whereas only 20% ($N=99$) of other sentences produced during the immediate recall task were produced verbatim after the delay. This difference is significant in a chi-square test (with Yates correction, $\chi^2(1,N=667) = 107.83, p < 0.01$). Thus subjects were more likely to repeat utterances that they had both heard and said than those phrases that they had only said earlier.

We also ran a logistic regression analysis to determine how predictive the two factors, hearing a phrase and saying a phrase, are on whether or not the delayed response is a verbatim match to the original story. Both factors were significant predictors, with coefficients of 3.17 ($p < 0.0001$) for having heard the clause and 2.19 ($p < 0.0001$) for having said the clause. Thus it seems that speakers are remembering what they had said during their own earlier recall protocol, and they are also remembering what they heard.

**Discussion**

Experiment 5 provides clear evidence for long-term verbatim memory, although it clearly does decay over time. Subjects were able to recall utterances that they had heard two days earlier, particularly if they also were able to recall them immediately. The reinforcing of things heard by things said is actually very natural in language acquisition (e.g., Lieven et al. 2003). So it is not as strange at it may seem at first to study recall in a context where heard and said utterances are intermixed.

**Conclusions and Implications**

We have seen that verbatim memory for language exists in naturalistic contexts, in which subjects are exposed to fairly long, non-“interactive,” coherent stories, when no warning of a memory task is given, and when possible lexical memory effects are controlled for. Significant verbatim memory has been demonstrated in recognition by above-chance performance. Verbatim memory has also been shown in recall, even after a two day delay.

The emphasis in early studies on the lack of verbatim memory for language was based on comparisons between memory for meaning and memory for structure. When compared directly, it is very clear that meaning always trumps structure. At the same time, we do appear to have an impressive ability to recognize and recall specific utterances verbatim. These findings lend credence to the possibility that language is learnable on the basis of categorizing over the input: possibly partially abstracted
utterance tokens are retained and may be categorized into more abstract constructional schemas (cf. e.g., Bybee, 1985; Tomasello, 2003; Goldberg, 2006).

The present studies only open the door to future research in this area. For example, the relationship between implicit and explicit memory for language deserves further attention. The two phenomena have rarely discussed in the same breath, and when they are, it is generally only to contrast them in passing. For example in a recent review article on (implicit) structural priming, Branigan (2007:12) writes, “a surprising feature of syntactic priming, given people’s poor explicit memory for surface form, is that it can be highly persistent…” One difference between verbatim memory and structural priming that might be suggested is that structural priming is abstract whereas verbatim memory needs to be lexically specific. Yet it is recognized that lexical specificity encourages structural priming. That is, priming is increased when lexical items are shared between prime and target (Branigan & Pickering, 1998; Cleland & Pickering, 2003; Savage et al., 2003). In fact, in comprehension tasks involving argument structure constructions, priming has so far only been found when the verbal head of the phrase is repeated (Arai, Gompel, & Scheepers, 2007; Branigan, Pickering, & McLean, 2005). The size of the effect of structural priming and verbatim memory are also comparable. Structural priming generally yields an increase of 10-20% in productions of the primed structure. Our experiments likewise reveal 10-20% of utterances are recalled verbatim. More direct comparisons between explicit recognition and recall and implicit measures could well be illuminative.

Other questions are also outstanding. Verbatim recall decays somewhat over time; how much time is required before it disappears altogether? Are certain types of constructions easier to recall or recognize than others? Factors such as construction type, lexical and construction frequencies, sentence length and complexity could well be relevant factors. Which specific models of memory can explain the detailed pattern of retention and decay that exists in both explicit and implicit memory for language?

A great deal of research within the “usage-based” model of language learning has been aimed at understanding when and why learners make the particular generalizations they do. The present studies provide the somewhat belated prerequisite finding that learners retain significant explicit, and remarkably detailed memories for language, over which generalizations and abstractions may be formed.
References

Story books

References:


Table 1

*Formulaic and non-formulaic utterances*

<table>
<thead>
<tr>
<th>A. Utterances recognized to be formulaic</th>
<th>B. Utterances recognized to be non-formulaic</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Once you’ve done that the rest is easy.</em></td>
<td><em>Once you’ve been there, the rest is easy.</em></td>
</tr>
<tr>
<td><em>I see what you mean.</em></td>
<td><em>I feel what you mean.</em></td>
</tr>
<tr>
<td><em>I’ll believe it when I see it.</em></td>
<td><em>When I see it, I’ll believe it.</em></td>
</tr>
<tr>
<td><em>It just goes to show, you can’t be too careful.</em></td>
<td><em>It goes to show you, you can be too careless.</em></td>
</tr>
<tr>
<td><em>You’re not allowed to do that.</em></td>
<td><em>You’re allowed to do that.</em></td>
</tr>
</tbody>
</table>
Table 2

*Spiderman: Mean probabilities of "yes" and "no" responses*

<table>
<thead>
<tr>
<th></th>
<th>Matching</th>
<th>Non-matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of &quot;yes&quot;</td>
<td><strong>0.86 (Hits)</strong></td>
<td>0.41 (False alarms)</td>
</tr>
<tr>
<td>Probability of &quot;no&quot;</td>
<td>0.14 (Misses)</td>
<td><strong>0.59 (Correct rejections)</strong></td>
</tr>
</tbody>
</table>

*Note.* Correct responses are indicated by boldface.
Table 3
Gramps and the Fire Dragon: Mean probabilities of “yes” and “no” responses

<table>
<thead>
<tr>
<th></th>
<th>Matching</th>
<th>Non-matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of &quot;yes&quot;</td>
<td><strong>0.81 (Hits)</strong></td>
<td>0.63 (False alarms)</td>
</tr>
<tr>
<td>Probability of &quot;no&quot;</td>
<td>0.19 (Misses)</td>
<td><strong>0.37 (Correct rejections)</strong></td>
</tr>
</tbody>
</table>

*Note.* Correct responses are indicated by boldface.
Table 4
Spiderman: Cumulative numbers of Correct Matches (to heard story) and Coincidental Matches (to alternative version of the story).

<table>
<thead>
<tr>
<th></th>
<th>Correct Matches</th>
<th>Coincidental Matches</th>
<th>Total # of clauses heard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Story version 1</td>
<td>96</td>
<td>49</td>
<td>430</td>
</tr>
<tr>
<td>Story version 2</td>
<td>48</td>
<td>5</td>
<td>430</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>54</td>
<td>860</td>
</tr>
</tbody>
</table>
Table 5
Gramps and the Fire Dragon: Cumulative numbers of Correct Matches (to heard story) and Coincidental Matches (to alternative version of the story)

<table>
<thead>
<tr>
<th>Story</th>
<th>Correct Matches</th>
<th>Coincidental Matches</th>
<th>Total # of clauses heard*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1</td>
<td>52</td>
<td>20</td>
<td>370</td>
</tr>
<tr>
<td>Version 2</td>
<td>33</td>
<td>7</td>
<td>370</td>
</tr>
<tr>
<td>Total</td>
<td>85</td>
<td>30</td>
<td>740</td>
</tr>
</tbody>
</table>

* Total number of clauses = (37 x 10 people)
Table 6
*Spiderman immediate and delayed recall results*

<table>
<thead>
<tr>
<th></th>
<th>Immediate</th>
<th></th>
<th></th>
<th>Delayed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Coincidental</td>
<td>Total # of sentences heard</td>
<td>Correct</td>
<td>Coincidental</td>
<td>Total # of sentences heard</td>
</tr>
<tr>
<td>Version 1</td>
<td>91</td>
<td>21</td>
<td>430</td>
<td>65</td>
<td>18</td>
<td>430</td>
</tr>
<tr>
<td>Version 2</td>
<td>60</td>
<td>5</td>
<td>430</td>
<td>43</td>
<td>7</td>
<td>430</td>
</tr>
<tr>
<td>Overall</td>
<td>151</td>
<td>26</td>
<td>860</td>
<td>108</td>
<td>25</td>
<td>860</td>
</tr>
</tbody>
</table>
Figure 1 - Spiderman: matches vs. coincidental matches
Figure 2 - Gramps: matches vs. coincidental matches
Figure 1
Matches vs. coincidental matches (average per subject)

Figure 2
APPENDIX A: COMPLETE STORY VERSIONS


Maybe you’ve heard of me. Maybe you’ve seen me. I am Spidey-Man.

I wasn’t always a superhero. I used to be a regular guy, just like everyone else. Well, not exactly like everyone else.

I was sort of a geek.

I really like school. But it wasn’t always easy for me. I didn’t always fit in.

Some of the kids didn’t like me.

Everything changed when we went on a field trip to learn about spiders. Man-made spiders. Super-spiders

I used to be skinny. I used to have to wear glasses. Not anymore. Not since that spider bit me.

Now I am fast.

Now I am strong.

I don’t like to fight, but sometimes I have no choice.

The kids at school were amazed at my new strength.

Like the spider that bit me, I have an extra sense.

My spider sense warns me when someone needs my help.

After the spider bit me, I began to make webbing. At first I couldn’t control it.

So I practiced. Now I can take aim and shoot my webbing wherever I want it to go.

I swore to use my powers to help the people of this city. But I had to protect myself. If the bad guys knew who I was, my family would be in danger.

And so I became Spider-Man.

Like a spider, I can climb walls.

My fingers can stick to anything.

I can swing through the city on a strand of webbing.
I can flip, dip, and whirl.

I am strong enough to fight four bad guys at once!

I can go places no one else can.

As Spider-Man, I am always on the lookout for evil. Sometimes it comes looking for me.

I’ll always be Peter Parker. But when people are in danger, I am Spider-Man.

**Spiderman Story Version #2**

You may know me. You may have seen me. Spiderman is my name.

I didn’t used to have superpowers. I was just like everyone else. Actually, I guess I wasn’t really ever just like everyone else.

I was called a geek.

School was interesting. But I had a hard time. Fitting in was the problem.

At school, I wasn’t liked by some of the kids.

One day, we took a field trip to learn about spiders. That day everything changed.

I used to be skinny and wear glasses. But not after that spider bit me.

I am super speedy now.

I’m also super strong.

I don’t like fighting, but I don’t have any choice sometimes.

My new strength amazed my classmates.

I have an extra sense just like the spider that bit me.

When my help is needed, I get a warning from my spider sense.

I started making webs after the spider bit me. I didn’t know how to do it at first.

After long hours of practice, I am now able to shoot my webbing wherever I aim it.
The people of this city needed me so I promised to help them. But if my identity were
known by the bad guys, I could put my family in danger.

So I became Spiderman.

Just as a spider can climb walls, so can I.

I can stick my fingers to anything.

A strand of webbing helps me swing through the city.

I can whirl and flip and dip.

Fighting four guys at once is easy for someone as strong as me.

No one else can go where I can go.

Spiderman is always on the lookout for evil. Sometimes I come face to face with it.

I’m still Peter Parker. But I’m Spiderman if anyone needs me!


It’s bedtime,” said Jesse.
“But I’m not sleepy.”
“Me neither,” said Gramps.
And they rocked in front of the fire, back and forth, back and forth.
“Look into the fire,” said Gramps.
“Can you see pictures?
I can see an old twisted apple tree.”
“Me, too,” said Jesse.
Together, they saw a flower garden, and a path leading to a castle.
And then –
-- they saw a fire dragon.
“Yipes!” cried Jesse.
The fire dragon roared and began to chase them,
down the path, through the flower garden, and up the tree!
Gramps tossed apples at the dragon.
“Have a snack,” he said.
But that only made the dragon angry.
Gramps and Jesse hitched a ride on a passing hot-air balloon.
Roaring, the dragon followed.
And they rocked in front of the fire, back and forth, back and forth.
“Look into the fire,” said Gramps.
Jesse and Gramps zipped down a high mountain.
The dragon zipped too.
They raced through a jungle.
The dragon raced, too.
They crossed a wide river,
The dragon crossed, too.
Then they ran down a long dark tunnel.
But the dragon ran, too.
The dragon was so close, his flames licked Jesse’s heels.
“Run for your life!” shouted Gramps.
The dragon moved towards Gramps.
I’ll save you, Gramps!” cried Jesse. Quickly, he flagged down a fire truck.
Jesse grabbed a hose and sprayed the dragon.
The dragon got smaller, and smaller, and smaller, until there was no dragon at all.
“Just embers left,” said Gramps, looking into the fire.
“I guess I saved you, then,” said Jesse.
“You sure did,” said Gramps.
Gramps yawned. Jesse yawned.
Then they rocked in front of the fire, back and forth, back and forth,
until they fell asleep, together.

Gramps Story Version #2

“Time for bed” said Jesse.
“But I’m not at all sleepy.”
“I’m not either” said Gramps.
And they rocked and rocked, back and forth, in front of the fireplace.
“Take a look at the fire” said Gramps.
“Are you able to see pictures?”
“I see a twisted old apple tree.”
“I do too” said Jesse.
Together, they could see a flower garden, and a path that led to a castle.
Just then--
--a fire dragon appeared.
Jesse cried, “Yipes!”
Roaring, the fire dragon started to chase them.
He chased them down the path and through the flower garden and up the tree.
Gramps tossed the dragon apples.
“Snack on these” he said.
But that only angered the dragon.
A hot air balloon passed, and Gramps and Jesse hitched a ride on it.
With a roar, the dragon followed them.
And they rocked and rocked, back and forth, in front of the fireplace.
“Take a look at the fire” said Gramps.
Jesse and Gramps went down a high mountain, zipping along.
The dragon went zipping along, too.
They went racing through a jungle.
The dragon went racing, too.
They went across a wide river,
The dragon went across, too.
Then they ran through a long dark tunnel.
But the dragon ran through the tunnel, too.
The dragon got so close, his flames licked at Jesse’s heels.
Gramps shouted, “Run for your life!”
The dragon approached Gramps.
Jesse shouted, “I’ll save you, Gramps!”
Jesse grabbed a hose. He sprayed the dragon.
The dragon got smaller, smaller, smaller…until there was no dragon at all.
“There are just embers left” said Gramps as he looked into the fire.
“I saved you, didn’t I?” said Jesse.
“You surely did” said Gramps.
Gramps yawned and Jesse yawned.
Then they rocked and rocked, back and forth, in front of the fire
Until they both fell asleep.