LANGUAGE, COGNITION AND NEUROSCIENCE, 2016
http://dx.doi.org/10.1080/23273798.2016.1245425

Routledge
Taylor & Francis Group

False memories for varying and non-varying words in American English

Anne Pycha

Department of Linguistics, University of Wisconsin, Milwaukee, Milwaukee, WI, USA

ABSTRACT

Certain types of variation are licenced by phonotactics. For example, the American English phoneme /t/ varies word-finally (*ba*[t] ~ *ba*[ʔ]), but not word-initially. We used a Deese–Roediger–McDermott false memory paradigm to pursue the hypothesis that, given comparable information in the speech stream, varying words (such as *bat*) activate less strongly than non-varying words (such as *tip*). We presented listeners with lists of phonological neighbours, such as *rat, ban, bet,* etc. (neighbours of *bat*), and *lip, tin, type,* etc. (neighbours of *tip*), followed by recall and recognition tasks. Results showed that participants often "remembered" the unheard words *bat* and *tip,* replicating previous work indicating that activation from neighbours can produce false memories. Most importantly, false memory rates were significantly lower for *bat* compared to *tip,* suggesting that the presence of phonotactic conditions for variability affected the lexical activation of a word, compared to the absence of such conditions.

ARTICLE HISTORY
Received 29 October 2015
Accepted 7 September 2016

KEYWORDS
Variation; reduction; false memory; lexical activation

Introduction

Any word will vary from one surface pronunciation to the next, depending upon who says it and in what context. But some words vary more, and differently, than others. In American English, for example, speakers can reduce coronal nasal-stop sequences to nasal flaps in post-stress position, *ce*[nt]*er* ~ *ce*[ɾ̃]*er* (Ranbom & Connine, 2007), delete schwa in word-medial position, *cal*[ə]*rie* ~ *calrie* (Connine, Ranbom, & Patterson, 2008; Pinnow & Connine, 2014), and reduce voiceless coronal stops to glottalised realisations in word-final position, *ba*[t] ~ *ba*[ʔt̚] ~ *ba*[ʔ] (Deelman & Connine, 2001; Sumner & Samuel, 2005). In each of these cases, the variation is not random, but is permitted or denied by the phonotactics of the word itself. Speakers do not, for example, reduce nasal-stop sequences in pre-stress position (*ce*[nt]*aur* ~ **ce*[ɾ̃]*aur*), delete schwa in word-final position (*choler*[ə] ~ **choler*), or reduce [t] in word-initial position ([t^{h}]*ip* ~ *[ʔ]*ip*).

Previous research has demonstrated that the presence of such variability modulates the activation of lexical representations. To begin with, reduced variants exact a processing cost. For example, Ranbom and Connine (2007) compared [nt] versus [ɾ̃] variants for words like *center* in lexical decision and goodness rating tasks; Connine et al. (2008) compared [ə]-retained versus [ə]-deleted variants for words like *calorie* in syllable-count and lexical decision tasks; and Sumner and Samuel (2005) compared [t], [ʔt], and [ʔ] variants for words like *flute* in semantically primed lexical decision, as well as repetition-primed lexical decision and yes/no recognition. A key finding from all of these studies was that listeners recognised unreduced variants more accurately and/or more quickly than reduced variants (see also Ernestus, Baayen, & Schreuder, 2002; Kemps, Ernestus, Schreuder, & Baayen, 2004). Previous research also demonstrates that variant recognition succeeds only in licenced phonotactic environments. For example, Pitt (2009) used the Ganong paradigm (Ganong, 1980) to show that listeners treated nonsense variants like [ˈsɛ.ɾ̃i] (based on real words like *twenty,* where nasal flapping is licenced because it occurs in a post-stress syllable) as if they were actual words, but crucially did not do so for stimuli like [ˈshɑ.ˌɾ̃ɛnt] (based on real words like *content,* where nasal flapping is not licenced because it occurs in a stressed syllable) (see also Brouwer, Mitterer, & Huettig, 2012, 2013; Gaskell & Marslen-Wilson, 1998; Mitterer & Blomert, 2003; Mitterer & McQueen, 2009). Taken together, these results provide support for an inference-based model of variant recognition, according to which listeners essentially "undo" the phonological rules that give rise to variation in order to access a single underlying citation form (Gaskell & Marslen-Wilson, 1998). In this model, (a) no inference is required for unreduced forms, which correctly predicts that such forms should exhibit an advantage, and (b) inference makes reference to a specific phonological environment, which correctly

CONTACT Anne Pycha pycha@uwm.edu
Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/23273798.2016.1245425

predicts that variant recognition does not succeed outside of this environment.

Despite this previous work, however, we still do not know how words that could vary – whether they actually do or not – differ from those that cannot vary. It is conceivable, for example, that the mere presence of the phonotactic conditions for variability may affect the lexical activation of a word, compared to the absence of such conditions. If so, we would predict differences in activation for varying versus non-varying words, such as *calorie* versus *cholera, center* versus *centaur,* and *bat* versus *tip* – even when those words occur in citation forms. To our knowledge, no previous work has investigated this prediction. But related research on frequency-sensitivity provides some reasons to think it will borne out. In a series of studies, Connine and colleagues have demonstrated that, in addition to the privileged status of unreduced variants, the most frequent variants of a particular word also facilitate lexical activation, regardless of whether those variants are reduced or unreduced. For example, the word *calorie* retains schwa with high frequency, while *broccoli* retains schwa with low frequency. In lexical decision to variants that retain schwa, listeners responded more quickly to *calorie* [kæləɹi] than to *broccoli* [bɹɑkəli]; in the same task with variants that delete schwa, listeners responded more slowly to *calorie* [kælɹi] than to *broccoli* [bɹɑkli] (Connine et al., 2008). Similar results held for words like *mental,* which retains [nt] with high frequency, versus *center,* which retains [nt] with low frequency (Ranbom & Connine, 2007).

These results are at odds with phonological inference models, which claim that only unreduced forms are represented in the lexicon, and suggest instead the need for a hybrid model in which both unreduced and reduced forms, as well as their accompanying frequency information, are represented in the lexicon. As proposed by Connine and colleagues, such a model makes predictions based upon frequency differences between individual lexical items, but we can apply a similar logic to frequency differences between entire classes of lexical items. For example, initial-/t/ words, such as *tip,* do not permit reduction to glottal stop and therefore retain the unreduced [t] variant with very high frequency. However, final-/t/ words, such as *bat,* do permit such reduction and therefore occur with [ʔ] in some tokens, but [t] in others. For the sake of simplification, we can state that initial-/t/ words occur with unreduced [t] roughly 100% of the time, while final-/t/ words occur with [ʔ] less than 100% of the time (for one way to quanitfy these numbers more precisely, see Crystal & House, 1988, p. 1557). Thus, if the most frequent variant of a word facilitates lexical activation, then a spoken variant with [t] should activate /t/-initial words more strongly than a spoken variant with [ʔ] activates /t/-final words. For example, all other things being equal, [tʰɪp] (100% frequency) should activate *tip* more strongly than [bæʔ] (less than 100% frequency) activates *bat*.

Although we have derived this prediction using a hybrid frequency-sensitive model, the inference model happens to make a similar prediction in this particular instance. This is because [tʰɪp] is an unreduced form that requires no "undoing" of phonological rules in order to activate *tip,* while [bæʔ] is a reduced form that requires inference to activate *bat,* which should lead to lower, or slower, activation. Interestingly, however, the logic of the hybrid model also applies when we consider unreduced tokens of varying words: while initial-/t/ words occur with [t] 100% of the time, final-/t/ words occur with [t] less than 100% of the time, so [tʰɪp] (100% frequency) should activate *tip* more strongly than [bæt] (less than 100% frequency) activates *bat*. Here, the competing models diverge. The hybrid model predicts differences in activation between words like *tip* versus *bat* – even when they occur in citation forms. The inference model, by contrast, treats all citation forms alike and therefore predicts equivalent levels of activation in this instance.

Under the scenario we have outlined, it does not particularly matter whether an individual word such as *bat* occurs with [t] more or less frequently than other words such as *dot* or *hate*. What matters is that initial-/t/ words, as a class, occur with [t] more frequently than final-/t/ words occur with either [t] or glottal stop, respectively. The broader implication would be that phonotactically licenced variability delineates whole classes of words whose representations differ from one another, such that comparable information in the speech stream activates non-varying words (*tip, take, tide,* etc.) more strongly than it activates varying words (*bat, dot, hate,* etc.).

Testing this prediction, however, presents a serious methodological challenge (for one approach, see Ranbom & Connine, 2007: Experiment 3). A key problem is that word-initial positions play a disproportionately important role in activating representations, quite independently of variation (e.g. Gaskell & Marslen-Wilson, 1997; Marslen-Wilson & Welsh, 1978 and many others). To take one concrete example from many available in the literature: although [sœro:χa:t] is the only Dutch word which begins with [sœro:] or ends with [o:χa:t], listeners were more likely to identify such words correctly if they heard the initial fragment, compared to the final fragment (Nooteboom, 1981). Thus, when listeners map the speech stream directly onto representations, as they do in most experimental tasks,

phoneme position within the word dramatically affects experimental outcomes. Unfortunately, word-initial privilege makes the same prediction that the hybrid model makes, namely that spoken [t] should play a bigger role in activating *tip* than it does in activating *bat*. This situation would seem to preclude a meaningful comparison between non-varying and varying words whose phonotactics differ solely in terms of word position.

To address this challenge, the current study uses the Deese–Roediger–McDermott (DRM) false memory paradigm (seminal papers include Deese, 1959; Roediger & McDermott, 1995; for reviews, see Gallo, 2006, 2010). In this paradigm, participants see or hear a list of words, such as *thread, pin, eye, sewing, sharp, point,* etc., that are semantic or phonological associates of a critical item, such as *needle*. On subsequent memory tasks, participants (falsely) remember seeing or hearing the critical item about 40–55% of the time, even though it was not present on the list (Roediger & McDermott, 1995). In addition to semantic associates, this phenomenon has also been reported for phonological neighbours: after hearing a list of words such as *rack, pack, bag, bat, book, bake,* etc., participants falsely remember the critical item *back* about 65–70% of the time (Ballardini, Yamashita, & Wallace, 2008; Garoff-Eaton, Kensinger, & Schacter, 2007; McDermott & Watson, 2001; Schacter, Verfaellie, & Anes, 1997; Sommers & Lewis, 1999; Wallace, Stewart, & Malone, 1995; Wallace, Stewart, Sherman, & Mellor, 1995; Watson, Balota, & Roediger, 2003). Dozens of additional studies have reported similar findings, both for free recall tasks in which participants write down the words they remember from a list, and for recognition tasks in which listeners hear a word and give a yes/no response to indicate if they remember it (Gallo, 2006, 2010). In our study, we asked participants to listen to lists of neighbours such as *rat, ban, bet,* etc. (for the critical item *bat*) versus *lip, tin, type,* etc. (for the critical item *tip*) and we asked whether rates of false recall and/or recognition differed for words with /t/ in final versus initial position.

Crucially, the DRM false memory paradigm reliably activates lexical representations in a manner that is insensitive to phoneme position within the word. The effect can be explained by the concept of converging neighbourhood activation (for a different view, see Kroll, Knight, Metcalfe, Wolf, & Tulving, 1996; Reinitz, 2001). Previous work has demonstrated that when a listener hears a signal such as [lɪp], she activates the representation for the target *lip,* but also the representations for phonological neighbours that differ in the substitution of one phoneme, such as that for *tip, lope, lid,* and so on (Luce & Pisoni, 1998). In the false memory paradigm, such activation occurs repeatedly, and eventually converges on the single word which is a neighbour to all of the items on the list, namely the critical item (e.g. *tip*), giving listeners the experience of thinking that they heard it, even though they did not (Collins & Loftus, 1975; Roediger, Balota, & Watson, 2001; Sommers & Lewis, 1999). In addition, the relative activation level of the critical item modulates the probability of false remembering (Robinson & Roediger, 1997; Wallace, Stewart, Sherman, et al., 1995; see also Roediger, Watson, McDermott, & Gallo, 2001). For example, neighbours such as *rack, pack, bag, bat, book, bake,* etc. are strongly associated to the word *back* and also produce high rates of false remembering for *back,* but neighbours such as *shack, yak, ban, batch, beak, bike,* etc. are only weakly associated to *back* and consequently produce lower rates of false remembering (Sommers & Lewis, 1999). We can therefore interpret rates of false recall and/or recognition as reflecting levels of spreading activation for the representation of the critical item.

Importantly for our purposes, the critical item in the DRM false memory paradigm receives activation from multiple sources, including neighbours that differ according to the phoneme which occurs in initial position, final position, and medial position. For example, the representation for critical item *tip* receives activation from spoken *lip* [lɪp] (initial C substitution), *tin* [t^{h}ɪn] (final C substitution), and *type* [t^{h}a^{j}p] (medial V substitution). Analogously, the representation for critical item *bat* receives activation from spoken *rat* [ɹæt] (initial C substitution), *ban* [bæn] (final C substitution), and *bet* [bɛt] (medial V substitution). Because activation arrives indirectly from diverse neighbours, rather than directly from the speech signal, initial segments should not disproportionately affect activation compared to final segments. Indeed, Westbury, Buchanan, and Brown (2002) verified this idea by comparing lists of phonological neighbours with initial CV overlap (*bade, bane, beige,* etc.), to lists with final VC overlap (*make, wake, sake*, etc.), and found no significant difference in false memory rates for critical items (*bake*). Thus, a key advantage of the false memory approach is that we can attribute any differences between false memories for *tip* versus *bat* to purely phonotactic differences, rather than to task-dependent asymmetries in the role played by [t] in initial position.

For the current study, one consequence of relying upon indirect activation is that we did not investigate the specific question of whether spoken stimuli with [t] are more likely to directly activate targets such as *tip* compared to *bat*. Instead, we investigated the related question of whether spoken stimuli such as [t^{h}ɪn] and [t^{h}a^{j}p] are more likely to indirectly activate the neighbour *tip,* compared to stimuli such as [ɹæt] and [bɛt] indirectly activating the neighbour *bat*. In doing so, we are

assuming a transitive relationship whereby relatively high or low activation levels on a target word spread those relatively high or low activation levels to their respective neighbours. That is, we assume that if [tʰɪn] produces relatively high activation for *tin,* it also produces relatively high activation for the neighbour *tip*. Concomitantly, if [ɹæt] produces relatively low activation for *rat,* it also produces relatively low activation for the neighbour *bat*. Support for this assumption comes from previous false memory experiments that have manipulated the relative activation of studied words, either by increasing repetitions from, for example, one to three (Benjamin, 2001; Hall & Kozloff, 1970; Underwood, 1965), or, for visual stimuli, by increasing duration of exposure from, for example, 20 to 2000 milliseconds (Arndt & Hirshman, 1998; Kawasaki & Yama, 2006; Zeelenberg, Plomp, & Raaijmakers, 2003). Results show that heightened activation of heard or seen words does indeed produce higher rates of false remembering for the associated critical item (for some caveats to this conclusion, see Gallo, 2006, pp. 120–125). Following the logic of these results, the current study assumes that if heightened activation occurs on studied words like *tin* and *type* (compared to *rat* and *bet*), this will produce higher rates of false remembering for the associated critical item *tip* (compared to *bat*).

To recap, in the two experiments described below, we asked participants to listen to lists of neighbours such as *rat, ban, bet,* etc. (for the critical item *bat*) versus *lip, tin, type,* etc. (for the critical item *tip*), and to complete recall and recognition tasks. We predicted weaker activation – reflected in lower rates of false remembering – for critical items like *bat* compared to *tip*. The two experiments were identical except for the realisation of /t/ in final position. In Experiment 1, list words were pronounced with glottal stop in final position (e.g. [ɹæʔ] for *rat* and [bɛʔ] for *bet*). In Experiment 2, list words were pronounced with released [t] in final position (e.g. [ɹæt] for *rat* and [bɛt] for *bet*). (In both experiments, list words were pronounced with [tʰ] in initial position (e.g. *tin* [tʰɪn] and *type* [tʰaʲp])). To preview, results supported our predictions, showing significantly lower rates of false remembering for /t/-final compared to /t/-initial critical items, for both recall and recognition tasks. This result occurred in Experiment 1, where both hybrid and inference models predicted a difference in activation levels. Importantly, this result also occurred in Experiment 2, where only the hybrid model predicted a difference.

Experiment 1: reduced tokens with [ʔ]

The purpose of both Experiments 1 and 2 was to determine whether rates of false recall and/or recognition differed for critical items that were /t/-initial (*tip*) or /t/-final (*bat*). The design of the experiments was identical, except for the surface realisation of /t/ in final position. The distinguishing feature of Experiment 1 was that all list words were pronounced with glottal stop in final position, for example, [ɹæʔ] for *rat*, [bɛʔ] for *bet,* [fæʔ] for *fat*, [ðæʔ] for *that*, and so on. Note that, in accordance with the rules of American English phonotactics, list words were pronounced with aspirated stop consonants in initial position, for example, [tʰɑp] for *top*, [tʰaʲp] for *type,* [tʰɪf] for *tiff,* and so on.

Method

Stimulus design

As critical items, we selected 36 mono-syllabic American English words with CVC structure. Nine words contained /t/ in initial position (*tip, toll, tide,* etc.), nine words contained /t/ in final position (*bat, wit, dot,* etc.), and eighteen words lacking /t/ (*ham, bid, lice,* etc.) served as fillers. Across the two conditions, we balanced critical items for frequency, familiarity, and phonological neighbourhood density to the extent possible, as displayed in Table 1.

As Table 1 shows, it was not possible to achieve a perfect balance, particularly for neighbourhood density. This is because the English lexicon does not contain large numbers of highly familiar mono-syllabic words with /t/, so the set of words that we drew from was small.

For each critical item, we constructed a list of 10 phonological neighbours that differed from it by the substitution of a single phoneme in initial, medial, or final position. Sample lists are displayed in Table 2, and the full set of lists is displayed in the Appendix.

As far as possible, we included roughly equal numbers of neighbours differing in C_1, V, or C_2 on each list, although the lexicon of English as well as the overall constraints of our false memory design limited the extent to which we could achieve this goal. Table 3 displays the mean proportion of neighbour types included on the lists. Since there are three neighbour types, a perfectly balanced set of lists would indicate 0.33 in each cell; as

Table 1. Lexical statistics (means and standard deviations) for the 18 words used as critical items in the false memory paradigm, across two conditions.

Condition	Log frequency	Familiarity	Density
Initial /t/	2.39 (0.66)	6.85 (0.23)	25.67 (3.32)
Final /t/	2.41 (0.66)	6.94 (0.09)	30.78 (6.00)

Notes: Log frequency is the base-10 log of the overall corpus frequency ("WU Speech & Hearing Lab Neighborhood Database Site," n.d.). Familiarity ratings represent judgments on a scale from 1 to 7, from a large sample of American English speakers (Nusbaum, Pisoni, & Davis, 1984). Density refers to the total number of words that differ from the item by the addition, deletion, or substitution of one phoneme ("WU Speech & Hearing Lab Neighborhood Database Site," n.d.).

Table 2. Sample lists of words for the false memory paradigm, each of which is a phonological neighbour of the indicated critical item.

Condition	Critical item	C_1 neighbours	V neighbours	C_2 neighbours
Initial /t/	*tip*	chip, hip, lip, rip, ship, whip	top, type	tiff, tin
Final /t/	*bat*	fat, rat, that	bet, boot, bout	badge, ban, bass, bath

Table 3. Mean proportion of neighbour types included on lists of 10 words, across conditions.

Condition	C_1 neighbours	V neighbours	C_2 neighbours
Initial /t/	0.57	0.21	0.22
Final /t/	0.30	0.38	0.32

Note: C_1 neighbours differ from the critical item in the initial consonant (e.g. *tip* vs. *chip*), V neighbours differ from it in the vowel (*tip* vs. *top*), and C_2 neighbours differ from it in the final consonant (*tip* vs. *tiff*).

is evident, the list deviates from this ideal, particularly in the Initial /t/ condition.

As with the critical items, we balanced the neighbours for frequency, familiarity, and phonological neighbourhood density to the extent possible, as displayed in Table 4.

Again, we did not achieve a perfect balance, both because of inherent limitations in the lexicon of English (the set of neighbours for a given critical item is fixed, and limited in number) and because of the need to roughly balance across neighbour types of C_1, V, and C_2.

Recording

A native speaker of English recorded each word in a sound-proof booth with a head-mounted microphone. The speaker was a phonetically trained male speaker of the Midwestern variety of American English, who was not aware of the purpose of the experiment. He recorded the words in a random order. The audio recording was digitised at a sampling rate of 44.1 kHz, and segmented into individual files using the Praat programme (Boersma & Weenink, 2014). For /t/-initial words, the speaker naturally produced a released and aspirated [t^h], in accordance with the phonological grammar of American English: for example, [t^h]*op,* [t^h]*ype,* [t^h]*iff*. For /t/-final words in Experiment 1, we asked the speaker to deliberately produce a glottal stop: for example, *fa*[ʔ], *ra*[ʔ], *tha*[ʔ]. The speaker did this consistently and with no difficulty.

Table 4. Lexical statistics (means and standard deviations) for the 180 words used as neighbours in the false memory paradigm, across two conditions.

Condition	Log frequency	Familiarity	Density
Initial /t/	2.20 (0.75)	6.51 (0.99)	23.91 (6.33)
Final /t/	2.38 (0.77)	6.73 (0.81)	25.68 (6.92)

The author, plus two additional American English speakers with formal training in phonetics, verified the presence of glottal stop (and absence of released [t]) in each of the /t/-final stimuli.

Procedure

The 36 lists were divided into 3 sets of 12, each containing 3 lists from the /t/-initial condition, 3 lists from the /t/-final condition, and 6 filler lists. Each participant was randomly assigned to one of the sets.

During the experiment, participants were seated in an individual carrel within a quiet laboratory setting, in front of a computer equipped with a mouse, keyboard, and high-quality headphones. Printed instructions on the computer screen guided them through each step. In the first phase of the experiment, participants listened to 12 lists of ten spoken words. Each word on a list was played individually, followed by 1 second of silence before the onset of the next word. As described above, the 10 words were all neighbours of a critical item, but crucially did not include the critical item itself. After each list, participants did a recall task, in which they were given 45 seconds to type as many words as they could remember from the list. After 45 seconds, they proceeded to the next list. The overall order of the 12 lists, as well as the order of the 10 words within each list, was randomised for each participant.

In the second phase of the experiment, after listening to all 12 lists, participants did a recognition task in which they listened to an individual spoken word, and made a yes/no judgment as to whether they had heard the word previously in the experiment. There were 96 items in the recognition task, which included 36 words that the participant actually heard (3 from each of the 12 heard lists), plus 60 that the participant had not heard. The unheard words included the 12 critical items from the participant's own set. In addition, the unheard words included 48 foils, consisting of 12 critical items from other experimental sets (one from each of 12 unheard lists), and 36 neighbour words from other sets (3 from each of 12 unheard lists). The order of items in the recognition task was randomised for each participant.

Participants

Participants were native speakers of the Midwest variety of American English ($n = 66$), between the ages of 18 and 30, approximately half female and half male, with no history of problems in speech, language, or hearing. The experiment took approximately 45 minutes of their time, and in exchange for participating, they received either cash compensation or extra credit points in a linguistics course.

Results: recall task

The recall task yielded a total of 5419 responses. Thus, on average, listeners typed 6.84 responses per each 10-word list (=5419/(12 lists * 66 participants)). After removal of filler items, a total of 3878 responses remained. Participants sometimes typed the same word twice for one list, resulting in 182 duplicates (91 in /t/-initial condition, 91 in /t/-final condition), which were removed from the data set. Participants also sometimes typed variant spellings, or mis-spellings, which were included in the data set as long as their orthographic-to-phonetic conversion produced a real English word. Thus, for example, *base* was accepted for *bass, tail* was accepted for *tale, teer* was accepted for *tier,* and *cote* was accepted for *coat*. There were 15 mis-spellings and typos which did not produce a real English word, such as *chlk* and *ig* (7 in /t/-initial condition, 8 in /t/-final condition), and these were removed from the data set. After removal of duplicates and mis-spellings, 3681 responses were included in the final analysis.

Following the procedures established in previous studies on false recall (e.g. Roediger & McDermott, 1995; Sommers & Lewis, 1999), we classified a response as "veridical" if it corresponded to a word that actually occurred on the list (e.g. *chip, hip, lip, rip, ship, whip*, etc.), "intrusion" if it did not occur on the list (e.g. random intrusions such as *child, coffee, table,* etc.), and "critical item" if it corresponded to the critical item (e.g. *tip*).

For descriptive statistics, we calculated proportions, again following previously established procedures. For veridical proportions, we divided the number of veridical responses per list by 10, which was the number of words that actually occurred on each list. For example, if a participant provided six veridical responses for a list, the proportion of veridical responses would be 0.60. For intrusion proportions, we divided the number of intrusion responses by the total number of responses per list. For example, if a participant provided two intrusion responses for a list, plus six veridical responses and one critical item, the proportion of intrusion responses would be $0.22 = 2/(2 \text{ intrusion} + 6 \text{ veridical} + 1 \text{ critical item})$. Finally, the critical item proportion was calculated as 0 if the participant did not respond with the critical item, and 1 if they did. (As described below, our statistical analysis accommodated the fact that this procedure counts veridical responses more than once). Table 5 displays the mean proportion of veridical, intrusion, and critical item responses given across the two conditions.

Table 5. Proportions of response types provided by participants (means, standard deviations) in recall task for Experiment 1.

	Veridical	Intrusion	Critical item
Initial /t/	0.52 (0.17)	0.15 (0.17)	0.40 (0.49)
Final /t/	0.50 (0.18)	0.18 (0.20)	0.16 (0.36)

For inferential statistics, we departed from previous literature in order to employ a mixed logit model, a decision motivated by the documented problems with the use of ANOVA for analysis of categorical outcome variables (Jaeger, 2008) and by the desire to include random effects. We ran a single model that included veridical, intrusion, and critical item responses. Following the procedures for logit models outlined in Jaeger (2008), we counted each response as a "success", and each possible lack of response as a "failure". For example, if a participant provided six veridical responses for a list, we counted six successes, plus four failures, corresponding to the four words on the list of ten that he or she did not recall. If a participant provided two intrusion responses, we counted two successes, and the remaining responses (veridical plus critical item) as failures. If a participant provided the critical item as a response, we counted one success, and zero failures.

We analysed these results in a mixed logit model, implemented with the glmer() function from the lme4 package in R, with predictor variables of response type (Veridical vs. Intrusion vs. Critical item) and position (Initial /t/ vs. Final /t/), and with random intercepts for both participants and items. Models that included random slopes for response type over participants and items, and position over participant, failed to converge. We used treatment coding. "Initial" served as the baseline for position, because previous studies on phonological false memories have typically used critical items of this type (i.e. non-varying), and we were interested in how "Final" would deviate from this baseline. "Critical item" served as the baseline for response type, which allowed us to make our crucial comparison between Initial /t/ versus Final /t/ specifically for recall of critical items.

Note that the counting procedure which we adopted from previous studies ultimately counts each veridical response twice: once as a "success" for veridical responses, and once as a "failure" for intrusion responses. The literature does not appear to have a standard solution to this issue: while some previous studies employing this counting procedure have excluded intrusion responses from their statistical analyses of recall (e.g. by conducting *t*-tests comparing exclusively veridical versus critical item responses; Roediger & McDermott, 1995), others have included intrusion responses (e.g. by conducting one *t*-test comparing veridical versus critical item responses, and another *t*-test comparing veridical versus instrusion responses; Sommers & Lewis, 1999). We took a different approach here. Because our model used treatment coding (rather than sum coding) with

Table 6. Results of mixed logit regression model for recall task in Experiment 1.

	Estimate	Std. error	*z* value	Pr(>|*z*|)
(Intercept)	−0.40	0.15	−2.70	6.91×10^{-3}*
Response (Veridical)	0.48	0.15	3.10	1.91×10^{-3}*
Response (Intrusion)	−1.29	0.17	−7.74	1.01×10^{-14}*
Position (Final)	−1.29	0.25	−5.22	1.82×10^{-7}*
Response (Veridical): Position (Final)	1.20	0.26	4.70	2.67×10^{-6}*
Response (Intrusion): Position (Final)	1.52	0.27	5.65	1.61×10^{-8}*

Note: *indicates a statistically significant result.

critical item responses as a baseline, it tested for simple effects (rather than main effects) for veridical responses compared to the critical item baseline, and also for intrusion responses compared to critical item baseline. Importantly, then, it did not perform a direct comparison between veridical and intrusion responses. Such a comparison is not needed to demonstrate a false memory effect, which occurs whenever the probability of a critical item response is significantly higher than that of a random intrusion response, and also helps avoid potential issues with the analysis of double-counted responses. Table 6 displays the results of the model.

The model indicates three simple effects. Response type exerted a simple effect for both levels of the predictor. The odds of a response increased in the Veridical condition compared to the Critical Item baseline, by a factor of approximately 1.62 ($=e^{0.48}$). The odds of a response decreased in the Intrusion condition compared to the Critical Item baseline, by a factor of approximately 0.28 ($=e^{-1.29}$). These effects replicate previous findings in the literature and indicate that the false memory paradigm produced the intended result: that is, participants were crucially less likely to respond with a random intrusion compared to a critical item. Position also exerted a simple effect. The odds of a response decreased in the Final /t/ condition compared to the Initial /t/ baseline, by a factor of approximately 0.28 ($=e^{-1.29}$). This effect demonstrates that participants were less likely to falsely recall a critical item like *bat,* compared to one like *tip*.

The model also indicates two interactions. Compared to the baseline, the odds of a response increased in the Veridical Final condition by a factor of approximately 3.32 ($=e^{1.20}$) and in the Intrusion Final condition by a factor of approximately 4.57 ($=e^{1.52}$).

Results: recognition task

In the recognition task, participants responded to three different types of items. "Veridical" items actually occurred on a list that the participant heard (e.g. *chip, hip, lip,* etc.). "Intrusion" items did not occur on any list that the participant heard (foils drawn from one of the two experimental sets that the participant was not assigned to, e.g. *comb, meat, sap,* etc.). "Critical items" were critical items for which the participant heard lists of neighbours (e.g. *tip*).

For descriptive statistics, following previous work, we calculated proportions of "yes" responses to each type of item. For veridical proportions, we divided the number of "yes" responses by the total number of items of this type that were presented. For example, participants responded to three veridical items from each of the lists they heard; if they responded "yes" to two of these, their veridical proportion for this list was 0.67 (=2/3). Participants responded to 4 intrusions from each of 12 lists that they did not hear; if they responded "yes" to one of these, their intrusion proportion was 0.25 (=1/4). Finally, participants responded to 1 critical item from each of 12 lists that they did hear; if they responded "yes" to it, their critical item proportion was 1 (=1/1). Table 7 displays the mean proportions of "yes" responses for the three item types across conditions.

For inferential statistics, again following the procedures for logit models outlined in Jaeger (2008), we counted each "yes" response as a success and each "no" response as a failure. As described for the recall task, we analysed these results in a mixed logit model with predictor variables of item type (Veridical vs. Intrusion vs. Critical item) and position (Initial /t/ vs. Final /t/), and with a random intercept for participants (note that for the recognition task, it did not make sense to include a random intercept for items, because a given item could count as veridical on one list, but as an intrusion for another list). Models with random slopes for response type and/or position over participants failed to converge. We used treatment coding such that "Critical Item" served as the baseline for response type and "Initial" served as the baseline for position. Table 8 displays the results of the model.

The model indicates three simple effects. Response type exerted a simple effect for both levels of the predictor. The odds of a "yes" response increased in the Veridical condition compared to the baseline, by a factor of approximately 1.86 ($=e^{0.62}$). The odds of a "yes" response decreased in the Intrusion condition compared to the baseline, by a factor of approximately 0.14 ($=e^{-1.98}$). These effects again indicate that the false memory paradigm produced the intended result. Position also exerted

Table 7. Proportions of "yes" responses (means, standard deviations) given in recognition task in Experiment 1.

	Veridical	Intrusion	Critical item
Initial /t/	0.77 (0.28)	0.22 (0.23)	0.65 (0.48)
Final /t/	0.73 (0.27)	0.25 (0.25)	0.47 (0.50)

Table 8. Results of mixed logit regression model for recognition task in Experiment 1.

	Estimate	Std. error	*z* value	Pr(>|*z*|)
(Intercept)	0.66	0.17	3.82	1.34×10^{-4}*
Response (Veridical)	0.62	0.19	3.33	8.63×10^{-4}*
Response (Intrusion)	−1.98	0.18	−11.28	$>2.00 \times 10^{-16}$*
Position (Final)	−0.78	0.22	−3.61	3.10×10^{-4}*
Response (Veridical): Position (Final)	0.57	0.26	2.20	0.03*
Response (Intrusion): Position (Final)	0.93	0.24	3.85	1.18×10^{-4}*

Note: *indicates a statistically significant result.

a simple effect. The odds of a "yes" response decreased in the Final /t/ condition compared to the baseline, by a factor of approximately 0.46 ($=e^{-0.78}$). This effect demonstrates that participants were significantly less likely to falsely recognise a critical item like *bat,* compared to one like *tip*.

The model also indicates two interactions. Compared to the baseline, the odds of a "yes" response increased in the Veridical Final condition by a factor of approximately 1.77 ($=e^{0.57}$) and in the Intrusion Final condition by a factor of approximately 2.53 ($=e^{0.93}$).

Summary and discussion for Experiment 1: reduced tokens with [ʔ]

Consistent with previous findings from the DRM paradigm, the false memory effect was robust in both recall and recognition tasks, as indicated by the fact that participants were significantly less likely to recall or recognise a random intrusion, compared to the critical item. The new finding from Experiment 1 is that the false memory effect was modulated by whether the critical item was a varying versus a non-varying word, as indicated by the fact that participants were significantly less to falsely recall and recognise a final-/t/ item such as *bat* compared to an initial-/t/ critical item such as *tip*. As discussed, false memory rates provide a mechanism for probing levels of lexical activation. Thus, we can interpret the results of Experiment 1 to indicate that, given equivalent amounts of speech input, *bat* words exhibited significantly lower activation levels than *tip* words. This conforms to our predictions.

These findings are consistent with both the hybrid frequency-sensitive model and the inference model of variant recognition, albeit for different reasons. The hybrid model predicts that words which licence phonotactic variability, such as *bat,* should activate less strongly than those words that prohibit such variability, such as *tip*. The inference model, on the other hand, predicts that tokens presented in reduced form (e.g. [ɹæʔ] for *rat*) should activate less strongly than tokens presented in unreduced form (e.g. [tʰɪn] for *tin*). In Experiment 1, all words that licenced variability in final-/t/ position were also presented in their reduced forms. Thus, we cannot know whether the difference between initial-/t/ and final-/t/ conditions arose from the mere potential for variability, or from the actual fact of surface variation. This issue is addressed in Experiment 2.

Experiment 2: unreduced tokens with [t]

Experiment 1 was identical to Experiment 2, except that all stimulus words with final /t/ were pronounced with a released [t]. Thus, when listeners were exposed to words such as *fat, rat, that*, etc., they heard the surface forms [fæt], [ɹæt], [ðæt], etc.

Method

The method for Experiment 2 was identical to that of Experiment 1. The same stimuli and experimental design were used. The only difference was that the speaker, who was the same person who produced the recordings in Experiment 1, deliberately produced a released [t] for all instances of /t/ in final position. The speaker did this consistently and with no difficulty. The author, plus two additional American English speakers with formal training in phonetics, verified the presence of released [t] in each of the /t/-final stimuli. Note that for words with /t/ in initial position, and for filler words, the same tokens as in Experiment 1 were used.

Participants

Participants were native speakers of the Midwest variety of American English ($n = 74$), between the ages of 18 and 30, approximately half female and half male, with no history of problems in speech, language, or hearing. None of them had participated in Experiment 1. The experiment took approximately 45 minutes of their time, and in exchange for participating, they received either cash compensation or extra credit points in a linguistics course.

Results: recall task

The recall task yielded a total of 6017 responses. Thus, on average, listeners typed 6.78 responses per each 10-word list (=6017/(12 lists * 74 participants)). After removal of filler items, a total of 3045 responses remained. Using the same procedure as described for Experiment 1, we removed 120 duplicates (62 in /t/-initial condition, 58 in /t/-final condition) and 23 mis-spellings/typos from the data set (12 in /t/-initial condition, 11 in /t/-final condition), such that 2902 responses were included in the final analysis.

Table 9. Proportions of response types provided by participants (means, standard deviations) in recall task for Experiment 2.

	Veridical	Intrusion	Critical item
Initial /t/	0.53 (0.17)	0.13 (0.16)	0.34 (0.48)
Final /t/	0.49 (0.16)	0.18 (0.17)	0.20 (0.40)

The responses were classified as veridical, intrusion, or critical item, and response proportions were calculated in the same manner as described for Experiment 1. These are displayed in Table 9.

As in Experiment 1, we analysed these results in a mixed logit model with predictor variables of response type (Veridical vs. Intrusion vs. Critical item) and position (Initial /t/ vs. Final /t/), and with random intercepts for participants (models that also included a random intercept for items failed to converge, while a separate model that include only a random intercept for critical items yielded results nearly identical to those displayed below). Models with random slopes for response type over participants and items, and position over participant, failed to converge. We used treatment coding such that "Critical Item" served as the baseline for response type and "Initial" served as the baseline for position. Table 10 displays the results of the model.

The model indicates three simple effects. Response type exerted a simple effect for both levels of the predictor. The odds of a response increased in the Veridical condition compared to the Critical Item baseline, by a factor of approximately 2.23 ($=e^{0.80}$). Meanwhile, the odds of a response decreased in the Intrusion condition compared to the Critical Item baseline, by a factor of approximately 0.32 ($=e^{-1.14}$). As for Experiment 1, these effects replicate previous findings in the literature and indicate that the false memory paradigm produced the intended result: that is, while participants were more likely to respond with a veridical word compared to a critical item, they were also less likely to respond with a random intrusion compared to a critical item. Position also exerted a simple effect. The odds of a response decreased in the Final /t/ condition compared to the Initial /t/ baseline, by a factor of approximately 0.47 ($=e^{-0.75}$). As for Experiment 1, this effect demonstrates that participants were significantly less likely to falsely recall a critical item like *bat,* compared to one like *tip*.

The model also indicates two interactions. Compared to the baseline, the odds of a response increased in the Veridical Final condition by a factor of approximately 1.79 ($=e^{0.58}$) and in the Intrusion Final condition by a factor of approximately 3.03 ($=e^{1.11}$).

Table 10. Results of mixed logit regression model for recall task in Experiment 2.

	Estimate	Std. error	*z* value	Pr(>|*z*|)
(Intercept)	−0.66	0.14	−4.58	4.60×10^{-6}*
Response (Veridical)	0.80	0.15	5.37	7.91×10^{-8}*
Response (Intrusion)	−1.14	0.16	−7.11	1.18×10^{-12}*
Position (Final)	−0.75	0.22	−3.39	6.92×10^{-4}*
Response (Veridical): Position (Final)	0.58	0.23	2.53	0.01*
Response (Intrusion): Position (Final)	1.11	0.24	4.60	4.19×10^{-6}*

Results: recognition task

For descriptive statistics, we followed the same procedure as in Experiment 1 for calculating the proportion of "yes" responses to each type of item. Table 11 displays the mean proportions for the three item types across conditions.

As in Experiment 1, we counted each "yes" response as a success and each "no" response as a failure and analysed the results in a mixed logit model with predictor variables of item type (Veridical vs. Intrusion vs. Critical item) and condition (Initial /t/ vs. Final /t/), and with random intercepts for participants. (As with Experiment 1, it did not make sense to include a random intercept for items, because a given item could count as veridical on one list, but as an intrusion for another list). Models with random slopes for response type and/or position over participants failed to converge. We used treatment coding such that "Critical Item" served as the baseline for response type and "Initial" served as the baseline for condition. Table 12 displays the results of the model.

The model indicates three simple effects and no interactions. Response type exerted a simple effect for both levels of the predictor. The odds of a "yes" response increased in the Veridical condition compared to the baseline, by a factor of approximately 2.61 ($=e^{0.96}$). The odds of a "yes" response decreased in the Intrusion condition compared to the baseline, by a factor of approximately 0.26 ($=e^{-1.36}$). These effects replicate previous

Table 11. Proportions of "yes" responses (means, standard deviations) given in recognition task, for Experiment 2.

	Veridical	Intrusion	Critical item
Initial /t/	0.78 (0.19)	0.29 (0.18)	0.58 (0.33)
Final /t/	0.71 (0.17)	0.23 (0.17)	0.47 (0.30)

Table 12. Results of mixed logit regression model for recognition task in Experiment 2.

	Estimate	Std. error	*z* value	Pr(>|*z*|)
(Intercept)	0.37	0.17	2.15	0.03*
Response (Veridical)	0.96	0.19	5.14	2.74×10^{-7}*
Response (Intrusion)	−1.36	0.17	−8.05	8.05×10^{-16}*
Position (Final)	−0.54	0.21	−2.56	0.01*
Response (Veridical): Position (Final)	0.18	0.25	0.70	0.49
Response (Intrusion): Position (Final)	0.21	0.23	0.92	0.36

findings in the literature and indicate that the false memory paradigm produced the intended result. Position also exerted a simple effect. The odds of a "yes" response decreased in the Final /t/ condition compared to the baseline, by a factor of approximately 0.58 ($=e^{-0.54}$). This effect demonstrates that participants were significantly less likely to falsely recognise a critical item like *bat,* compared to one like *tip*. No other effects reached significance.

Summary and discussion for Experiment 2: unreduced tokens with [t]

Replicating previous studies in the DRM paradigm, the false memory effect was robust in both recall and recognition, as indicated by the fact that participants were significantly less likely to remember random intrusions, compared to the critical item. Furthermore, and importantly for our hypothesis, the false memory effect was modulated by whether the critical item was a varying versus a non-varying word, as indicated by the fact that participants were significantly less likely to falsely remember a final-/t/ critical item such as *bat*, compared to an initial-/t/ critical item such as *tip*. Since false memory rates reflect levels of lexical activation, we interpret these results to indicate that, given equivalent amounts of speech input, words like *bat* exhibited significantly lower activation levels than words like *tip*. Thus, the results of Experiment 2, in which all words were pronounced in unreduced form, are very similar to those of Experiment 1, in which some words were pronounced in unreduced form while others were pronounced in reduced form.

These findings are consistent with the hybrid frequency-sensitive model of variant recognition, but not with the inference model. The hybrid model predicts that any word which cannot vary will activate less strongly than any word which can vary – regardless of whether a particular token occurs in a reduced form or not. Experiment 2 exhibits precisely this pattern of results. All words were presented in their unreduced citation forms (e.g. [tʰɪn] for *tin,* [ɹæt] for *rat*); despite this, results showed that final-/t/ words still exhibited significantly weaker activation than initial-/t/ words. The inference model, by contrast, predicts that only those particular tokens which occur in reduced form will activate less strongly than tokens which occur in unreduced form, and therefore cannot account for these results.

Summary and discussion

The current study was motivated by the question of how variation modulates the activation of lexical representations. Specifically, we were interested in whether words that licence variation by virtue of their phonotactic environment activate less strongly than words that do not licence such variation. Following the logic of a hybrid frequency-sensitive model of variant recognition (Connine et al., 2008; Ranbom & Connine, 2007), we hypothesised that words which do licence variation, and are therefore realised with a given variant less than 100% of the time, should activate less strongly than those words which do not licence such variation, and are therefore realised with a given variant approximately 100% of the time. Importantly, the hybrid model predicts that differences between non-varying and varying words should be present regardless of whether any given token occurs in its unreduced versus reduced form. That is, the mere potential for variation leads to reduced activation. This distinguishes it from the inference model, which predicts differences only between reduced tokens versus unreduced citation tokens.

To test our hypothesis, we focused on American English words such as *bat,* which licence variation such that word-final /t/ may be pronounced with either [t] or [ʔ], and compared them to words such as *tip,* which do not licence such variation. We used a DRM false memory paradigm, which provides a measure of the indirect activation that occurs after listeners have heard a list of phonological neighbours which converge on an unheard critical item. We hypothesised that spoken stimuli with final /t/, such as *rat* and *bet,* should produce relatively weak activation on their targets (*rat, bet*) and therefore on their unheard neighbours (*bat*) as well. Crucially, this relatively weak activation should occur both when final /t/ was pronounced in reduced form [ʔ] and also when it was pronounced in unreduced citation form [t]. Meanwhile, we hypothesised that spoken stimuli with initial /t/, such as *tin* and *type,* should produce relatively strong activation on targets (*tin, type*) and on unheard neighbours (*tip*).

The results supported this hypothesis. After hearing a list of phonological neighbours, participants were significantly less likely to falsely remember varying words such as *bat,* compared to non-varying words such as *tip*. This result occurred in both recall and recognition tasks for Experiment 1, where /t/-final neighbours were pronounced with reduced [ʔ]. Importantly, this result also occurred in both recall and recognition tasks for Experiment 2, where /t/-final neighbours were pronounced with unreduced [t].

The overall pattern of data appears highly similar across Experiments 1 and 2, as suggested by the recall data in Tables 5 and 9, and the recognition data in Tables 7 and 11. To verify this observation, we conducted a logistic regression analysis of the pooled data from

both experiments. Results showed that for the recall task, the factor of reduced [ʔ] (as in Experiment 1) versus unreduced [t] (as in Experiment 2) did not produce a simple effect ($\beta = 0.25$, std. error = 0.20, $z = 1.23$, $p = .22$), nor did it produce interactions. For the recognition task, this factor did not produce a simple effect ($\beta = 0.26$, std. error = 0.21, $z = 1.24$, $p = .22$), although it did participate in a three-way interaction. Compared to the baseline, the odds of a response in the Intrusion Final Reduced condition increased by a factor of approximately 2.01 ($\beta = 0.70$, std. error = 0.32, $z = 2.17$, $p < .05$), which suggests a difference in the pattern of intrusion responses. Although the origin of this difference is not clear, it may arise from the fact that the participants in Experiment 1 were different from those in Experiment 2. The crucial point is that, across experiments, there is no significant difference in patterns of responses to critical items. Thus, false memory rates for *tip* and *bat* in Experiment 1 are comparable to those in Experiment 2. These findings support the basic prediction derived from the hybrid model, namely that comparable information in the speech stream activates non-varying words more strongly than varying words. There are, however, a number of caveats to this conclusion, which we turn to in the following sections.

Activation of reduced versus unreduced tokens

The equivalent pattern of results across Experiments 1 and 2 is at odds with previous studies that have reported an advantage for unreduced citation forms (Connine et al., 2008; Ranbom & Connine, 2007). For example, Sumner and Samuel (2005) reported that unreduced tokens of words like *flute* [flut] primed subsequent presentations of the same word in both lexical decision and yes/no recognition tasks, but reduced tokens like [fluʔt] and [fluʔ] did not. The implication for the current study is that false memory rates for words like *bat* should have been somewhat higher in Experiment 2 (due to relatively higher activation produced by the unreduced tokens like *rat* [ɹæt] and *bet* [bɛt]) compared to those in Experiment 1 (due to relatively lower activation produced by the reduced tokens like *rat* [ɹæʔ] and *bet* [bɛʔ]). Indeed, provided that false memory rates for *bat* had remained lower than those for *tip* in both experiments, such a result would have still supported our predictions and also been more fully compatible with the hybrid model, which is a "hybrid" precisely because it incorporates both an advantage for unreduced citation forms as well as information about frequency of occurrence.

Nevertheless, the lack of a significant difference between false memory rates for critical items in Experiments 1 versus 2 should be interpreted with caution, both because it is a negative result and also because it was not an explicit goal of the current study to compare unreduced versus reduced variants of the same word. Future work, incorporating modified experimental designs as well as multiple experimental methods for probing activation, will need to address this issue.

Other consonants in initial versus final position

One limitation of the current study is that we examined only a single type of variation in a single language, that is, reduction to glottal stop in word-final position in American English words. Future work will need to determine whether the current findings apply to other sets of varying versus non-varying words, such as *calorie* versus *cholera,* and *center* versus *centaur,* and to other languages.

Another limitation is that we do not know if the current results are due specifically to the variability versus non-variability of /t/, or more generally to the final versus initial positioning of any consonant. To address this issue, we would need to compare critical items such as *dish* [dɪʃ] versus *chef* [ʃɛf], or *dip* [dɪp] versus *pan* [pæn], where a target consonant other than /t/ (i.e. /ʃ/ or /p/, respectively) occurs in different word positions. Previous work has shown that consonants such as /ʃ/ do not exhibit significant positional variations in pronunciation. For example, in an electropalatography study, the sibilants /s, z, ʃ, ʒ/ showed equivalent patterns of peak contact in initial versus final positions (Keating, Wright, & Zhang, 1999). Meanwhile, consonants such as /p/ are susceptible to small contextual changes in final position, even if they do not vary as significantly as /t/. For example, in a corpus of naturally produced American English, labial /p/ was realised as a complete stop 23% of the time in final position, but 88% of the time in initial position (Crystal & House, 1988, p. 1557). Given these findings, we might speculate that false memory rates would be generally lower for /C/-final versus /C/-initial critical items, but that the differences for consonants like /p/ and especially /ʃ/ would nevertheless be smaller than those for /t/-final versus /t/-initial critical items. For now, given this limitation of the current study, the strongest claim we can make is that the *potential* for variation – by virtue of occupying final position, rather than initial – modulates rates of false memories and, by extension, activation of lexical representations.

Role of lexical characteristics

Another potential limitation of the current study comes from the lexical characteristics of the neighbours used

as stimuli. Due to gaps in the English lexicon, it was not possible to perfectly match the frequency, familiarity, and density of the neighbours across the three experimental conditions, as Table 4 showed. As the discussion below reveals, however, it is highly unlikely that the modest differences in neighbour characteristics affected the key conclusions of our study – if anything, our results are potentially more robust because they occurred despite these differences.

The mean log frequency of neighbours in the Final condition (2.38) was higher than that in the Initial condition (2.20). Differences in frequency arguably affect activation levels of lexical representations (this issue is debated; see Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998), so this asymmetry would lead us to believe that critical items such as *bat* should exhibit higher rates of false memories than those such as *tip*, if high-frequency words cause more spreading activation to neighbours than low-frequency words do. But our results revealed the opposite pattern: *bat* exhibited lower rates of false memories than *tip*, strongly suggesting that a separate factor was at play. If frequency had any role in our pattern of results, it served only to reduce the size of the difference between Final versus Initial conditions.

The lexical characteristics of neighbour words also exhibited modest differences in mean familiarity of neighbours, which was higher in the Final condition (6.73, on a scale of 1–7) than that in the Initial condition (6.51). It is difficult to formulate specific predictions about how these differences could have impacted our results because, to our knowledge, no theory links these specific characteristics to activation. The false memory literature does use the concept of "familiarity" to explain certain phenomena (Yonelinas, 2002) although we are not aware of any previous work that employs scaled word familiarity ratings of the kind we report here.

Neighbour words also exhibited modest differences in mean neighbourhood density, which was greater in Final (25.68) versus Initial (23.91) condition. High-density words by definition have more neighbours than low-density words, and in line with this, the psycholinguistics literature predicts that hearing a high-density word should activate a larger number of unheard neighbours than hearing a low-density word (Goldinger et al., 1989; Luce & Pisoni, 1998). But to our knowledge, the literature does not predict a difference in activation on any individual neighbour. For example, unheard *bat* can be activated by heard neighbour *rat,* which has a relatively high density of 37. And, unheard *tip* can be activated by heard neighbour *whip,* which has a relatively low density of 10. We do not, on this basis, necessarily predict any difference in activation of *bat* versus *tip;* we only predict a difference in the number of other words that are simultaneously activated. This question warrants further investigation, particularly in light of previous work which has demonstrated that, under certain conditions, words with equivalent neighbourhood densities can nevertheless produce different patterns of overall activation (Chan & Vitevitch, 2009; Vitevitch, 2002, 2007; Vitevitch & Luce, 1999). For now, we tentatively conclude that modest differences in density across conditions probably did not confound our results.

Finally, our stimuli also exhibited differences in the number of neighbour types across the three conditions. Our goal was to include roughly equal numbers of C_1, V, and C_2 neighbours in each condition (e.g. *lip, type, tin,* respectively, as neighbours of *tip*), but constraints on the English lexicon prevented this. The biggest difference is that, as depicted in Table 3, the Initial condition contained a comparatively large proportion of C_1 neighbours (mean = 0.57), compared to that of the Final condition (mean = 0.30). If one type of neighbour increases the likelihood of a false memory more than other types of neighbours, then this difference across conditions could have played a confounding role in producing our results. As mentioned in the Introduction, however, Westbury et al. (2002) found no significant difference in false memory rates for critical items produced by lists containing only C_2 neighbours compared to those containing only C_1 neighbours, suggesting that our modest differences in neighbour types across conditions do not account for our results.

Direct versus indirect activation

Whereas most experimental paradigms directly activate lexical representations via matching segments in the speech stream, the DRM false memory paradigm is unique because it indirectly activates lexical representations via phonological neighbours. There are several implications to this, and we discuss three of them here.

First, we did not ultimately evaluate the specific idea that spoken stimuli with /t/ should directly activate /t/-final words less strongly than /t/-initial words. This would have required us to present participants with matching segments in the speech stream, such as [t^hɪp] versus [bæt] or [bæʔ], and measure the direct activation of *tip* versus *bat*. As discussed in the Introduction, the well-established asymmetry between initial and final segments in lexical activation precludes such a direct comparison. Instead, we employed indirect activation to investigate a more general prediction, namely that comparable information – here, in the form of matched lists of phonological neighbours – activates varying

words less strongly than it activates non-varying words. One disadvantage of this approach is that our results are not wholly comparable to the results of direct activation experiments, such as those of Connine and colleagues. Another disadvantage is that we must assume a transitive relationship whereby weaker (or stronger) activation for a target also produces weaker (or stronger) activation for its neighbours, although, as discussed in the Introduction, this assumption is supported by the results of previous studies that provide evidence for such a relationship (Arndt & Hirshman, 1998; Benjamin, 2001; Hall & Kozloff, 1970; Kawasaki & Yama, 2006; Underwood, 1965; Zeelenberg et al., 2003).

Second, the DRM false memory paradigm differs from more common tasks because, in addition to lexical activation, it includes a monitoring component. That is, the paradigm probes for participant responses to critical items that, while partially activated in the mental lexicon, were never actually heard. As a result, participants must decide whether their memory for a particular word is real – that is, they must monitor the source of their memories before providing a response (Johnson, 2006; Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981; Roediger & McDermott, 1995). Thus, false memory rates for two words which are equivalently activated can nevertheless differ, if the monitoring component evaluates those two words differently. For example, Schacter and colleagues (Israel & Schacter, 1997; Schacter, Israel, & Racine, 1999) showed that participants who studied words accompanied by pictures (e.g. *thread* accompanied by a picture of thread) exhibited significantly lower false recognition rates than participants who studied words without pictures. A subsequent study showed that participants who said words aloud at study (e.g. they saw the printed word *thread* and also said "thread" out loud) exhibited significantly lower false recognition rates than participants who only saw the words (Dodson & Schacter, 2001). Source-monitoring can account for these findings: while the unheard critical item (e.g. the semantic associate *needle*) was presumably activated for both sets of participants, those in the aloud condition did not remember saying the word, and therefore rejected the activation of the critical item. In other words, they monitored the source of their memories using the heuristic "if I do not remember saying the word *needle* out loud, I probably did not see the word *needle,* either"; on this basis, participants could either accept the memory as "real" (even though it was not) or reject it.

Given this previous work, we must remain open to the idea that differences in monitoring, rather than in activation, could account for our results. For example, it is conceivable that lists of phonological neighbours created equivalent activation for words like *tip* and *bat,* but that the subsequent monitoring process led participants to reject a higher number of memories for *bat,* compared to *tip*. Such a scenario remains highly speculative, however, given that no obvious monitoring heuristic (i.e. something akin to "if I did not say the word aloud, I must not have seen it on the list, either") occurred in our experiments.

Third, although we have interpreted our results primarily in terms of activation for different types of representations, they also have implications for the phenomenon of false remembering more generally. The experience of remembering an event that did not actually occur is not uncommon; as Loftus and Bernstein (2005, p. 11) state: "One does not have to look far to find compelling cases in which individuals have held distorted memories about events from their past" (see also Loftus, 2005). Furthermore, false memories of the kind we examine here, that is, for unheard or unseen neighbours or associates, are remarkable for being robust and highly replicable (Gallo, 2006, 2010). Indeed, such false memories can occur even when the list of neighbours is quite small (Robinson & Roediger, 1997), or when (orthographic) neighbours are presented for only brief durations (Seamon, Luo, & Gallo, 1998), and they typically induce a strong subjective belief that the critical item was truly heard or seen (Gallo, 2006, Chapter 4). Given the tenaciousness of the effect, the results of the current study are notable for identifying one simple way to reduce it, namely, by introducing the potential for phonotactically licenced variation. If low false memory rates for *bat* words eventually generalise to other types of varying words, such as *calorie* ~ *calrie* and *center* ~ *cenner,* we could potentially conceive of the lexicon as being stratified according to the relative susceptibility of, or resistance to, individual representations to distortions of memory.

Conclusion

Although much work remains to be done, the current study, to our knowledge, provides the first evidence that the very presence of a licencing environment for variation can affect the lexical activation of a word, compared to the absence of such conditions. That is, lexical activation for a varying word like *bat* is weaker than for a non-varying word like *tip* – even when tokens occur in unreduced forms. Thus, phonotactically licenced variability appears to delineate whole classes of words whose representations differ from one another, with consequences not just for activation, but also for subsequent remembering.

Acknowledgments

I thank Ellen Abolt, Rebecca Lehr, Dylan Pearson, and Amara Sankhagowit for research assistance. I thank Ivan Ascher, Fred Eckman, and Teri Mannes for assistance with pilot tests. I also thank audiences at the Midwestern Phonology Conference (2014), and the International Congress of Phonetic Sciences (2015) for feedback, and Eleni Pinnow and an anonymous reviewer for suggestions that greatly improved the manuscript. Any flaws are mine alone.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Research Growth Initiative, University of Wisconsin, Milwaukee [101X292].

References

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*(3), 371–391. doi:10.1006/jmla.1998.2581

Ballardini, N., Yamashita, J. A., & Wallace, W. P. (2008). Presentation duration and false recall for semantic and phonological associates. *Consciousness and Cognition*, *17*(1), 64–71. doi:10.1016/j.concog.2007.01.008

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 941–947. doi:10.1037/0278-7393.27.4.941

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer (Version 5.4.04)*. Retrieved from www.praat.org

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539–571. doi:10.1080/01690965.2011.555268

Brouwer, S., Mitterer, H., & Huettig, F. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics*, *34*(3), 519–539. doi:10.1017/S0142716411000853

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1934–1949. doi:10.1037/a0016902

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. doi:10.1037/0033-295X.82.6.407

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception & Psychophysics*, *70*(3), 403–411. doi:10.3758/PP.70.3.403

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America*, *83*(4), 1553–1573. doi:10.1121/1.395911

Deelman, T., & Connine, C. M. (2001). Missing information in spoken word recognition: Nonreleased stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(3), 656–663. doi:10.1037/0096-1523.27.3.656

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. doi:10.1037/h0046671

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, *8*(1), 155–161. doi:10.3758/BF03196152

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, *81*(1), 162–173. doi:10.1006/brln.2001.2514

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

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848. doi:10.3758/MC.38.7.833

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125. doi:10.1037/0096-1523.6.1.110

Garoff-Eaton, R. J., Kensinger, E. A., & Schacter, D. L. (2007). The neural correlates of conceptual and perceptual false recognition. *Learning & Memory*, *14*(10), 684–692. doi:10.1101/lm.695707

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*(5–6), 613–656. doi:10.1080/016909697386646

Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 380–396. doi:10.1037/0096-1523.24.2.380

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, *28*(5), 501–518. doi:10.1016/0749-596X(89)90009-0

Hall, J. W., & Kozloff, E. E. (1970). False recognitions as a function of umber of presentations. *The American Journal of Psychology*, *83*(2), 272–279. doi:10.2307/1421331

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review*, *4*(4), 577–581. doi:10.3758/BF03214352

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Johnson, M. K. (2006). Memory and reality. *American Psychologist*, *61*(8), 760–771. doi:10.1037/0003-066X.61.8.760

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. doi:10.1037/0033-2909.114.1.3

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. doi:10.1037/0033-295X.88.1.67

Kawasaki, Y., & Yama, H. (2006). The difference between implicit and explicit associative processes at study in creating false memory in the DRM paradigm. *Memory*, *14*(1), 68–78. doi:10.1080/09658210444000520

Keating, P., Wright, R., & Zhang, J. (1999). Word-level asymmetries in consonant articulation. In *UCLA working papers in phonetics* (pp. 157–173). Retrieved from https://www.researchgate.net/profile/Richard_Wright2/publication/244528405_Word-Level_Asymmetries_in_Consonant_Articulation/links/53d7dc270cf2631430bfc91f.pdf

Kemps, R., Ernestus, M., Schreuder, R., & Baayen, H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, *90*(1), 117–127. doi:10.1016/S0093-934X(03)00425-5

Kroll, N. E., Knight, R. T., Metcalfe, J., Wolf, E. S., & Tulving, E. (1996). Cohesion failure as a source of memory illusions. *Journal of Memory and Language*, *35*(2), 176–196. doi:10.1006/jmla.1996.0010

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–366. doi:10.1101/lm.94705

Loftus, E. F., & Bernstein, D. M. (2005). Rich false memories: The royal road to success. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 101–113). Washington, DC: American Psychological Association Press. Retrieved from https://webfiles.uci.edu/eloftus/LoftusBernsteinInHealy05.pdf

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. doi:10.1016/0010-0285(78)90018-X

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language*, *45*(1), 160–176. doi:10.1006/jmla.2000.2771

Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception & Psychophysics*, *65*(6), 956–969. doi:10.3758/BF03194826

Mitterer, H., & McQueen, J. M. (2009). Processing reduced word-forms in speech perception using probabilistic knowledge about speech production. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 244–263. doi:10.1037/a0012730

Nooteboom, S. G. (1981). Lexical retrieval from fragments of spoken words: Beginnings vs. endings. *Journal of Phonetics*, *9*(4), 407–424. Retrieved from http://alexandria.tue.nl/repository/freearticles/734601.pdf

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report*, *10*(10), 357–376. Retrieved from http://s3.amazonaws.com/academia.edu.documents/2328101/Hoosier_mental_lexicon_Nusbaum.pdf?AWSAccessKeyId = AKIAJ56TQJRTWSMTNPEA&Expires = 1475073604&Signature = LKY9NutDm0YNjJVcEBp4oPQ%2FkHQ%3D&response-content-disposition = inline%3B%20filename%3DSizing_up_the_Hoosier_Mental_Lexicon_Mea.pdf

Pinnow, E., & Connine, C. M. (2014). Phonological variant recognition: Representations and rules. *Language and Speech*, *57*(1), 42–67. doi:10.1177/0023830913479105

Pitt, M. A. (2009). How are pronunciation variants of spoken words recognized? A test of generalization to newly learned words. *Journal of Memory and Language*, *61*(1), 19–36. doi:10.1016/j.jml.2009.02.005

Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, *57*(2), 273–298. doi:10.1016/j.jml.2007.04.001

Reinitz, M. T. (2001). Illusions of memory. *Comments on Theoretical Biology*, *6*, 411–430. Retrieved from http://www.researchgate.net/profile/Mark_Reinitz/publication/232614345_Illusions_of_memory/links/09e4150880be751535000000.pdf

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*(3), 231–237. doi:10.1111/j.1467-9280.1997.tb00417.x

Roediger, H. L., Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association Press.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. doi:10.1037/0278-7393.21.4.803

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. doi:10.3758/BF03196177

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language*, *40*(1), 1–24. doi:10.1006/jmla.1998.2611

Schacter, D. L., Verfaellie, M., & Anes, M. D. (1997). Illusory memories in amnesic patients: Conceptual and perceptual false recognition. *Neuropsychology*, *11*(3), 331–342. doi:10.1037/0894-4105.11.3.331

Seamon, J. G., Luo, C. R., & Gallo, D. A. (1998). Creating false memories of words with or without recognition of list items: Evidence for nonconscious processes. *Psychological Science*, *9*(1), 20–26. doi:10.1111/1467-9280.00004

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, *40*(1), 83–108. doi:10.1006/jmla.1998.2614

Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language*, *52*(3), 322–338. doi:10.1016/j.jml.2004.11.004

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, *70*(1), 122–129. doi:10.1037/h0022014

Vitevitch, M. S. (2002). Influence of onset density on spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 270–278. doi:10.1037/0096-1523.28.2.270

Vitevitch, M. S. (2007). The spread of the phonological neighborhood influences spoken word recognition. *Memory & Cognition*, *35*(1), 166–175. doi:10.3758/BF03195952

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408. doi:10.1006/jmla.1998.2618

Wallace, W. P., Stewart, M. T., & Malone, C. P. (1995). Recognition memory errors produced by implicit activation of word candidates during the processing of spoken words. *Journal of Memory and Language*, *34*(4), 417–439. doi:10.1006/jmla.1995.1019

Wallace, W. P., Stewart, M. T., Sherman, H. L., & Mellor, M. D. (1995). False positives in recognition memory produced by cohort activation. *Cognition*, *55*(1), 85–113. doi:10.1016/0010-0277(94)00646-3

Watson, J. M., Balota, D. A., & Roediger III, H. L. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language*, *49*(1), 95–118. doi:10.1016/S0749-596X(03)00019-6

Westbury, C., Buchanan, L., & Brown, N. R. (2002). Sounds of the neighborhood: False memories and the structure of the phonological lexicon. *Journal of Memory and Language*, *46*(3), 622–651. doi:10.1006/jmla.2001.2821

WU Speech & Hearing Lab Neighborhood Database Site. (n.d.). Retrieved October 29, 2015, from http://neighborhoodsearch.wustl.edu/Home.asp

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864

Zeelenberg, R., Plomp, G., & Raaijmakers, J. G. (2003). Can false memories be created through nonconscious processes? *Consciousness and Cognition*, *12*(3), 403–412. doi:10.1016/S1053-8100(03)00021-7