Speech errors and the phonological similarity effect in short-term memory: evidence suggesting a common locus.

Mike P. A. Page, Alison Madge

University of Hertfordshire, UK.

Nick Cumming

University of Plymouth, UK

and

Dennis G. Norris

MRC Cognition and Brain Sciences, Cambridge, UK.

Abstract

In three experiments, we tested the hypothesis that those errors in immediate serial recall (ISR) that are attributable to phonological confusability share a locus with segmental errors in normal speech production. In the first two experiments, speech errors were elicited in the repeated paced reading of six-letter lists. The errors mirrored the phonological confusions seen in ISR. In a third experiment, participants performed ISR for four-word lists. Some of the lists were designed to encourage the exchange of onset consonants between adjacent words. ISR was shown to be sensitive to this manipulation, further supporting the common-locus hypothesis. The results are discussed in the context of theories of serial recall and of speech production, and are further related to neuropsychological data.

Introduction

In this article, we present evidence that the errors underlying the phonological similarity effect (PSE) in immediate serial recall (ISR) are similar in character to errors seen in spontaneous speech production. We therefore explore the hypothesis that the two types of error result from the operation of a common mechanism. The PSE is characterized by increased errors in the immediate serial recall of lists containing words that sound similar. This effect has been shown in lists of rhyming letters (e.g. "BGTCVP" Baddeley, 1968; Conrad, 1964; Conrad & Hull, 1964; Wickelgren, 1965), and with lists of rhyming words and/or words that share a vowel (e.g., Baddeley, 1966). More recent work (Fallon, Groves & Tehan, 1999; Nimmo & Roodenrys, 2004) has confirmed that order-memory for lists of CVC words suffers whether items share either rime (vowel and coda), or onset and coda, or onset and vowel. These studies also found that *item* recall can actually improve when items share a rime, probably because the rime is a salient cue that can assist in determining which items were present (though not the order in which they occurred). This enhanced item-memory can mask the PSE when a conventional correct-in-position scoring method is used, though not when order errors are conditionalized on a free-recall measure of the item in question. Nimmo and Roodenrys noted this sensitivity to the rime unit, and more generally the increased influence that vowel similarity exhibited in their data relative to similarity based on shared consonants. They related these factors to models, such as that of Hartley and Houghton (1996), that seek to place linguistic constraints on the representations used in both short-term memory and in speech production. Nimmo and Roodenrys concluded that there was “an urgent need for STM researchers to integrate linguistic research, and models based on this research, into STM models”. This paper is part of an attempt to do just that.

Other authors have drawn attention to the similarity between verbal STM and speech production. Most notably Ellis (1980), influenced by the earlier contributions of Morton (1964, 1968, 1970), explored the proposal that immediate recall of verbal materials was carried out using a “response buffer”. The primary function of this buffer was taken to be the storage of a speech programme during the period between between speech planning and overt articulation. Ellis proposed what he called the “error equivalence hypothesis”, namely, that if a common response buffer was involved in both speech production and short-term memory for serial order, then similar types of phonemic error would be expected in both tasks. In three experiments, involving recall of lists of CV and VC syllables, he corroborated this hypothesis by showing: that serial recall errors most often involved consonant swaps (more so than either vowel swaps or whole-syllable swaps); that consonant errors respected a feature-similarity effect, such that consonants tended to exchange when they were more featurally similar; that phonemic errors tend to preserve within-syllable position; and, finally, that consonant swaps were more numerous between syllables that shared a vowel, a pattern he dubbed the “contextual similarity effect”. All of these serial recall effects had previously been observed in relation to speech production errors (e.g., Mackay, 1970; Nooteboom, 1967).

Although it is not something to which Ellis (1980) drew direct attention, it is the contextual similarity effect that can be applied most directly to the PSE.

In a typical phonologically confusable stimulus-list of rhyming single-syllable items, the context similarity effect might be expected to apply with force: there are many onset consonants that share the same context, and this context is not just similar but is identical in both vowel and coda. Moreover, in rhyming lists, any onset-consonant exchange will result in the same items as in the stimulus list, just placed in a different order. Such exchanges are difficult for the speaker to detect because no unintended item is thereby introduced into their recall. The key observation is therefore that, although the extra order errors seen in recall of a list of rhyming items have traditionally been seen as exchanges of complete items, they might perhaps better be thought of as onset exchanges promoted by contextual similarity.

Explaining the PSE in terms of speech production mechanisms has a clear advantage from the perspective of computational models of verbal ISR. In the last decade or so, we (Henson, Norris, Page, & Baddeley, 1996; Norris, Page, & Baddeley, 1994; Page & Norris, 1998a; 1998b) and others (Burgess & Hitch, 1992; 1999; Henson, 1998) have used data from ISR tasks to help develop computational simulations of short-term memory. One of the most important constraints on these models was provided by data from recall of lists of alternating confusability. Such lists, including those in which rhyming items are placed at alternating list-positions (e.g., "BRPXDQ"), are interesting because it has been shown that while the rhyming items are subject to additional recall errors (usually mutual exchanges), the interleaved nonrhyming items are recalled as well as they would be in a list comprised entirely of nonrhyming items. Indeed, Farrell & Lewandowsky, 2003, have recently claimed that nonconfusable items are recalled *better* in mixed lists than in pure nonconfusable lists. For the alternating lists, this pattern of errors results in a serial position curve that has a characteristic saw-tooth shape, with error-peaks located at the stimulus-list positions occupied by confusable items. These data are difficult to explain in terms of "chaining" models of ISR, in which each list-item is associatively chained to its predecessor(s) (Henson et al., 1996). Moreover, early position-item association models (e.g., Burgess & Hitch, 1992) were unable to simulate such a pattern. The key factor that enabled the primacy model (Norris et al., 1994; Henson, et al. 1996, Page & Norris, 1998), and other later models (Burgess & Hitch, 1999; Henson, 1998) to simulate these data accurately was the incorporation of a two-stage, or two-phase, recall process. The precise details differ between models (see Page & Henson, 2001, for a review), but they all involve an initial stage/phase in which an item is selected on the basis of order information, followed by a second stage/phase, in which the selected item can be replaced at output by one with which it is phonologically confusable. Although the use of a two-stage mechanism allowed the models to simulate the PSE data, there appeared to be no independent motivation for using two stages. The second stage of the primacy model and related models does nothing other than introduce additional errors that would not occur in its absence. While the data seemed to require a second stage, its presence was thus something of an embarrassment. The appeal to parallels between ISR and speech production provides a way out of this somewhat uncomfortable situation.

In Page and Norris (1998a), and more explicitly in a companion chapter (Page & Norris, 1998b), we pointed out that most modern models of speech production are also inherently two-stage in nature. The requirement for two stages in models of speech production follows from the fact that speech is more than just a concatenation of discrete words. For a multiword utterance, once the phonological representation of the words is read from the lexicon, further processes must operate on those representations to produce speech that is fluent. A common assumption (e.g., Dell, 1986, 1988; Levelt, Roelofs, & Meyer, 1999) is that the production process engages a segmental plan, typically encompassing several upcoming words, and then fills the slots in this segmental plan using the corresponding set of phonemes that are primed in parallel. This two-stage process is deemed necessary to deal with those characteristics of continuous speech (e.g., resyllabification, porosody) that distinguish it from a simple concatenation of citation forms.

According to this line of reasoning, immediate serial recall of verbal materials, that is, speech *reproduction*, is highly related to more normal speech *production* in that both require the conversion from an ordered plan, specified at a somewhat abstract level, to fluent speech. The most parsimonious assumption is that both processes make use of much of the same machinery. This, in turn, offers an explanation for why ISR is performed using an error-prone and apparently redundant output stage. The errors do nothing to help ISR per se, but they are a consequence of using an existing mechanism that is itself optimized for the production of fluent speech.

Naturally, there are still issues to be resolved. One concerns the obvious disparity in the rate of second-stage errors for ISR (as, by hypothesis, seen in the PSE) and of speech errors in spontaneous speech. Speech errors are far less frequent than the order errors in ISR that result from phonological confusability. For a pure-confusable list of around (nonconfusable) span-length, one would expect to see at least one second-stage error in each list. Speech errors at such a rate would render speech virtually unintelligible. In Page and Norris (1998b) we offered several reasons why speech errors might be so much more common in the ISR of lists containing, for example, rhyming items. First, we noted that typical speech output is endogenous rather than exogenous: the words that a speaker intends to utter are generated internally by that speaker in the context of an intention to convey a particular meaning; as long as that meaning is kept in mind, then the speech planning process can be deliberate in its planning of the upcoming utterance. By contrast in the ISR task, the utterance that must be generated is required to be a repetition of some speech just heard (or some visual input just recoded): As noted above, the task is one of speech reproduction rather than just speech production. In this sense, the planned utterance is merely self-supporting and, in the view of many though not all STM theorists, decays over time. (It does not matter for our argument here if one assumes that the planned utterance is particularly vulnerable to interference, as opposed to its decaying). This fact, we suggested, might well make (exogenous) speech reproduction more prone to error than (endogenous) speech production.

A second and related point is that output and perforce rehearsal, in the context of the ISR task, are speeded in a way that is usually not necessary for the output of spontaneous speech. A speaker who finds themselves making speech errors in spontaneous speech has the option to slow their speech rate. But a participant trying to rehearse as much as possible in the gaps between the presentation of list items, does not have that luxury. Repeated speeded re-renderings of (part of) the list might plausibly raise the error rate.

Third, there is the fact, alluded to above, that it is very rare in spontaneous speech for a speaker to be required to utter six consecutive words, all of which share a rime. (Note that in the case of the rhyming letters B, C, D, etc., that are often used in tests of the PSE, the letter-names also share a syllable shape, a factor that Stemberger, 1990, has identified as further promoting speech errors.) As noted above, the shared rime in pure-confusable lists for ISR might not only promote the occurrence of speech errors, but will also render their detection extremely difficult. Unlike most speech errors, onset exchanges between rhyming items in ISR would introduce no new item, nor any change in "meaning".

In Page and Norris (1998b) we described detailed simulations of the relevant ISR data using a Dell-style (Dell, 1986, 1988) speech production model as the output stage of our primacy model (Page & Norris, 1998a). In doing so, we believed that this linking of models across two traditionally separate domains (pace Ellis and Morton) would be mutually beneficial. In the case of speech production models, it has not been customary to simulate the production of multiword utterances of a length approximating those used in immediate serial recall experiments. There appears to be general agreement, however, that "the activation of the plan causes anticipatory activation of units for upcoming elements" (Dell, Burger, & Svec, 1997, p. 128). This priming of upcoming words, with earlier words more active, is exactly the critical order-storage mechanism, the primacy gradient, around which our primacy model is built. Given that we have always identified the primacy gradient with Baddeley and Hitch's phonological loop (Baddeley & Hitch, 1974; Baddeley, 1986), this raises the possibility, as presaged by Ellis (1980), that the plan of a multiword utterance, and the phonological loop, are identical, at least in the special circumstances in which speech reproduction (or the reproduction of recoded visual information) is required. We will leave discussion of this matter until after presentation of our data, and will also address then some of the vexed questions regarding data from neuropsychological patients that arise as a consequence.

The data that are presented here follow on quite naturally from Ellis's (1980) experiments, with the focus now shifted to consideration of the PSE. The first two experiments examine a fairly straightforward question: in the speeded reading (rather than the remembering) of lists of alternating confusability, such as those that have been used in ISR tasks, do participants make speech errors that resemble the "memory" errors that are seen in the PSE? This is, in a sense, the reverse of Ellis's procedure: he asked participants to recall lists and noted that their errors resembled those known to occur in speech production; here, we ask people to speak (not remember) lists, to see whether their errors resemble those known to occur in recall from STM.

Experiment 1

As indicated above, in Experiment 1 we asked participants to perform one of two tasks. The first task involved the presentation of a series of lists of visually presented letters for immediate serial recall. These included four different types of list: pure nonconfusable; pure confusable; alternating confusability beginning with a nonconfusable; and alternating confusability beginning with a confusable. Confusability here was operationalized by using letters with rhyming letter-names. The second task involved participants reading such letter-lists aloud. Each letter-list was read ten times, each reading following on directly from the previous, at a speed indicated visually to the subject and thought to be sufficient to promote the occurrence of speech errors.

Method

Participants.

There were 40 participants, drawn from the University of Hertfordshire undergraduates. Their mean age was 20.3 years and, where relevant, they were offered course credit for their participation.

Materials.

The materials comprised 64 lists of 6 letters each. The letters were drawn from a pool including the rhyming letters B, C, D, G, P, T, and V, together with the nonrhyming letters H, J, L, Q, R, Y, and Z (pronounced 'zed' in British English). The first four lists were practice lists, each comprising 6 letters chosen randomly without replacement from the whole set. The experimental lists that followed were constrained in various ways: lists contained no repeated letter and no list was repeated; no letter followed its immediate alphabetic predecessor; no letter occupied the same position as it had in the previous list; no letter triplet appeared in consecutive lists; no list contained any obvious acronym; items were approximately matched for the number of occurrences in different list positions; and lists in all four conditions were matched separately for mean log bigram frequency (all $\log f$ in the range 4.8–5.0). Consistent with these constraints, fifteen lists were generated for each of the four list types: pure confusable (e.g., CTBDPV); pure nonconfusable (e.g., ZLJHRQ); alternating beginning confusable (e.g., TLVQCR); and alternating beginning nonconfusable (e.g., RDLCJB). Lists of different types were randomly intermixed to prevent, as far as was possible, any pattern's being detected. Finally, to avoid any spurious effects emerging due to a particular set of lists, eight different sets of 64 trials were generated consistent with the above constraints - four sets were used in the memory task and four in the speech production task. List-set was therefore a between-participant factor, with five participants randomly assigned to each set for each task.

Procedure.

Half the subjects performed an immediate serial recall task. Letters were presented visually one at a time at an inter-onset interval (IOI) of 750ms. Each letter stayed on the screen for 500ms, and was followed by a blank screen for 250ms. Letters were presented in black on a white background, in the centre of a computer screen, in a large font (approx. 2cm high). The computer screen was placed approximately 50cm from the participant. After the final letter had been presented, the word "Recall" appeared on the screen, where it remained for two seconds. After recalling the list, participants were asked to press the spacebar to begin the next trial, the first letter of which appeared one second later. Participants were asked to view the lists in silence and, when cued, to recall the list by writing it on a response sheet provided. Each line of the response sheet had 6 boxes and participants were required to write their recall in a strictly left-to-right fashion, indicating, with a dash in the appropriate box, any omissions in their recall. The experimenter was present and ensured compliance with these instructions. Previous responses were covered with a piece of paper to prevent their interfering with the current response. This part of the experiment took approximately 30 minutes.

The remaining half of the subjects performed a speech-error elicitation task. We used a paced repeated-reading task, similar to that adopted by Wilshire (1998, 1999), though at a somewhat higher speech-rate. Each of the lists in turn appeared in the centre of the computer screen. The entire list remained on the screen until the next trial was cued. Reading-rate was indicated by the change in color, from black to red, of a single letter in turn, at a rate of one change every 300ms. This ensured that participants continued looking at the to-be-read list throughout. A trial began with the list appearing on the screen in a large black font with no spaces between letters. 1500ms after the list appeared, the first letter turned red and remained red for 300ms before turning black again, immediately after which the next letter turned red, and so forth. The colour change cycled around the letters, such that the first letter in the list changed to red immediately after the final letter had reverted from red to black. This cyclic colour-change continued until the spacebar was pressed. Participants were asked to read the list at the pace indicated by the progress of the red letter. They were required to start reading the list ten times. If they made an error on any list, or if they stumbled, they were encouraged to complete the list and to start reading again on the next available list-initial colour-change. When ten attempts had been made for a given list, the experimenter hit the spacebar to start the next trial. All readings were recorded for later scoring, using a Marantz PMD650 professional minidisc recorder and studio-quality microphone.

Results

Immediate serial recall.

The ISR data were scored such that an item had to be recalled in the correct position to count as correct. The serial position curves for errors are shown in Figure 1. From inspection, the data appear consistent with previous work in the area: the fewest errors were in the pure nonconfusable condition; there was an increased number of errors in only the confusable positions of the alternating lists, resulting in a pronounced saw-tooth shape to the relevant serial position curves; the most errors were observed in the pure confusable condition. We derived "composite" confusable and nonconfusable data from the alternating-confusability conditions such that, for example, the composite confusable data comprised the error scores from odd positions in the alternating-beginning-confusable lists and from even positions in the alternating-beginning-nonconfusable lists. The data from the composite curves were subjected to a 4 (list-set: between) by 4 (confusability condition: within) by 6 (serial position) mixed-factor ANOVA. There were main effects of confusability condition, $F(3, 48) = 28.3, p < .001$, and serial position, $F(5, 80) = 9.0, p =< .001$, with no reliable main effect of list-set and no reliable interactions. Planned comparisons revealed that pure and composite nonconfusables did not differ, $t(16) = 1.55, p = .14$, but that both differed from both pure and composite confusables (all $t > 4.7, p < .001$) ; this is in line with expectations based on previous research (Baddeley, 1968; Henson et al. 1996). However, pure and composite confusables did not differ either, $t(16) = 1.8, p = .094$, which is not in line with that previous work though doesn't seriously contradict it. Nevertheless, the general data pattern is as expected, with a clear PSE and a clear saw-tooth shape to the mixed-list serial position curves.

Farrell and Lewandowsky (2003) found that, with appropriate controls on ensemble size, the two nonconfusable scores (pure and composite) did differ, with the composite nonconfusables recalled better than those in pure nonconfusable lists. There was a tendency for the same to be true in the current data though, as noted above, this difference was not reliable. Farrell and Lewandowsky's theoretical account of their finding predicted that the difference would be located in order errors. Looking at the order-accuracy measure (Fallon et al. 1999; Nimmo & Roodenrys, 2004), calculated as the number of times a given list-item was recalled in the correct position divided by the number of trials on which that item was recalled in any position, the tendency towards a difference between the pure and derived nonconfusable conditions weakens further (mean proportion order-accuracy .833 and .847 respectively, $t(16) = 1.06, p = .31$). If there is a qualitative difference between our result and Farrell and Lewandowsky's (which involves asserting a null result on our part) then it may have resulted from a particular design choice in their experiment. Lists in Farrell and Lewandowsky's experiments were blocked by list-type. They not only used lists of alternating confusability in their experiments (of which they only used the variety with nonconfusables in positions 2, 4 and 6), but also used lists in which a single nonconfusable was presented in a list that otherwise comprised confusables. Having blocked such lists, there was a possibility that participants became aware of the pattern (e.g., "the item that doesn't rhyme is always in the second position"). If this were the case, such abstract knowledge would enable participants to filter out possible order errors from their responses, leading to the small observed advantage for nonrhyming items in mixed lists versus those in pure lists. The question as to whether Farrell and Lewandowsky's result stems from the development of such abstract knowledge of list structure certainly warrants further investigation.

Speech-error elicitation.

To score the performance on the speech error-elicitation, we adopted the following procedure. First, we identified ten, and only ten attempts that that the participant made to read the list. These constituted the first ten readings that commenced at the list start. In these ten lists, we scored only the first error; this was to avoid the complications of dealing with subsequent attempts at correction, or subsequent comment by the participant. The serial position of each error was noted and the error itself was classified as detailed below.

The serial-position curves for speech errors are shown in Figure 2, They show the number of list attempts in which the first error occurred at the relevant serial position, taken as a proportion of the number of attempts that reached that serial position without prior error. What is immediate obvious is that there were more errors for the pure confusable lists than for the other lists types, and that there is a sawtooth pattern of errors in the alternating conditions, reminiscent of the pattern seen in the ISR data. These impressions were confirmed by the results of a 4 (list-set: between) by 4 (confusability conditions: within) by 6 (serial positions: within) mixed-factor ANOVA applied to the data corresponding to the composite curves derived as for the ISR task. There were main effects of confusability, $F(3, 48) = 34.6, p < .001$, and serial position, $F(5, 80) = 25.2, p < .001$, but none of list-set, $F(3, 16) = 1.75, p = .20$. In addition, there was a reliable interaction between condition and serial position, $F(15, 240) = 10.8, p < .001$, reflecting, it appears, the rather larger effect of condition in the earlier portions of the list. Both this interaction and the overall increase in errors for the first serial position in particular, are very likely a consequence of the demands of the repeated-reading task. As the colored letter cycled repeatedly through the stimulus list, the reading of the first item was nearly always (apart from on the first reading) immediately preceded by a (speeded) shift of gaze from the end of the list to the start. It is, we believe, reasonable to assume that this impacted negatively on the paced reading of the first item. Planned t-tests comparing confusability conditions indicated that all conditions differed from all others (all $t(16) > 3.4$, ps $\leq .004$), except for the pure and the composite nonconfusables, that were statistically indistinguishable, $t(16) = 0.37, p = .71$. This pattern, and the significant main effect of confusability, remained even when the data from the first serial position was removed from the analysis. The critical pattern seen in the ISR data is thus mirrored in the speech-error data.

Table 1 shows the classification of speech errors for the different types of list. (These figures do not include the small numbers of occasions on which an intruded letter or a nonletter response was given.) For any given list-type, each value has been scaled by the number of opportunities for an error of that type. For example, in an alternating list there are three possible confusable with confusable (C-C) substitutions but 9 possible confusable with nonconfusable (C-N) substitutions. The raw numbers of each substitution-type and the scaling factor are both given in brackets; for C-N substitutions in pure lists, for which the scaling factor would be zero, the raw number of such exchanges is given alone. The classification shows clearly that the additional errors on lists containing confusable items stem overwhelmingly from one confusable item's being read in place of another. It is also worth noting that of the 267 errors involving C-N substitutions (208 in mixed lists, 59 in pure lists), 118 can be accounted for mutual substitutions of the letters G and J (71 in mixed lists, 47 in pure lists). Obviously this particular error can also be seen as being promoted by phonemic overlap, though in the onset rather than in the rime.

Discussion

We have shown that the familiar pattern of increased errors in the ISR of confusable items in pure and mixed lists is mirrored in the speech errors that occur during the paced and moderately speeded reading of those lists. The increased errors were, in both cases, the result of increased substitutions of one confusable item for another confusable items or, perhaps, equivalent substitutions involving the onset consonants of their names. There was no ostensible memory component to the reading task, nor was there any primacy advantage that might indicate that the participants were trying to remember the lists rather than reading them. Because the reading procedure required participants to look at the stimulus list throughout, and because errors in the reading task are so much rarer than errors in the memory task, we feel strengthened in our inference that participants were not simply trying to remember the lists.

If we take it that the errors in the reading task are what would normally be classified as speech errors — they were, after all, produced in a task that has been traditionally and explicitly used for speech-error elicitation — then it appears that additional serial recall errors seen in lists containing confusable items share many of the characteristics of those speech errors, with the sole exception that the rate of commission of such errors is increased. We have already suggested reasons why there should be fewer errors in the speech production task, the most germane here being that for the production task, but not the memory task, the model for the intended utterance remains clearly visible to the participant. This "fixes" the utterance in a manner not dissimilar from the way in which we have suggested that an endogenous intention might.

These results are consistent with the hypothesis that there is a common underlying mechanism for certain types of speech error and those memory errors that underlie the phonological similarity effect in immediate serial recall. If this hypothesis is correct then the assumption of a second stage in various prominent models of ISR would be somewhat vindicated by relating them to models of speech production for which, as detailed above, the existence of two stages is better motivated.

Experiment 2

Given the theoretical import of the speech-error results of Experiment 1, we decided to replicate them in a second experiment. Rather than conducting an exact replication, we introduced a second factor that we expected to be independent of the confusability manipulation, namely irrelevant sound (IS). The experiment followed exactly the same procedure as the speech-elicitation task from Experiment 1, but in this case half the lists were accompanied by background irrelevant speech and half by white noise. It is a well established fact that irrelevant speech hinders the immediate serial recall of verbal materials (Colle & Welsh, 1976, etc.). Our own account of the underlying mechanism (Page & Norris, 2003) suggests that irrelevant speech (and other changing-state irrelevant sound; Jones, Madden, & Miles, 1992) can reduce the resources available for representing a to-be-remembered list as a primacy gradient of activations. Given that the proposal developed here suggests that the the primacy gradient might also be involved in other types of speech production (see General Discussion), we thought it would be interesting to investigate whether irrelevant speech could also have an effect on the speech-error elicitation task. The IS manipulation was necessarily rather exploratory, since the presence of an effect of irrelevant sound on speech-error commission would necessarily be rather more informative than its absence, but nonetheless we felt it worth including as part of a more general attempt to replicate the key findings of Experiment 1.

Participants.

There were 24 participants, drawn from the University of Hertfordshire undergraduates. Their mean age was 21.9 years and, where relevant, they were offered course credit for their participation.

Materials and Procedure.

Experiment 2 involved only the speech-error elicitation task. New list-materials were generated using the same constraints as for Experiment 1. The task and the marking procedure were also the same as those used for that experiment. Stimuli were presented in two blocks, with half the participants experiencing the first block in the presence of irrelevant speech and the second block in the presence of white noise; the remaining participants experienced these conditions in the reversed order. We used as the irrelevant speech, a passage of Finnish speech generated by digitally splicing together twelve six-second clips (16-bit, 22050 Hz sample rate) that had been used in previous IS experiments. This passage was played (over headphones) in a continuous loop throughout the relevant block. In the other block, white noise was presented in the same way, at the same subjective volume.

Results

The speech-error data were analyzed in a similar manner to that used in Experiment 1. A 4 (list-set: between) by 2 (block orders: between) by 2 (irrelevant speech/noise: within) by 4 (confusability conditions: within) by 6 (serial positions: within) mixed-factor ANOVA was applied to the data, for which composite nonconfusable and confusable curves had been derived from the alternating lists (as before). There were main effects of confusability, $F(3,45) = 19.8, p < .001$, and serial position, $F(5,75) = 19.9, p < .001$, but none of irrelevant sound ($F < 1$) or of the two between-participant factors block-order ($F < 1$) and list-set,$F(3,15) = 1.08, p = .39$. Of the interactions, only that between confusability and serial position was statistically significant, $F(15,225) = 4.0, p < .001$. As with the first experiment, the main effect of serial position and its interaction with confusability condition reflected a multiplicative increase in errors of all types at the first position. We again attributed this to the demands of the repeated-reading task.

Planned t-tests comparing confusability conditions indicated once again that all conditions differed from all others, all $t(15) > 3.0, p \leq .009$, except for the pure and the composite nonconfusables that were statistically indistinguishable, $t(15) = .83, p = .42$. Once again, both the significant main effect of confusability, and this pattern seen in the paired comparisons, were maintained in an analysis that omitted data from the first serial position. The original (noncomposite) serial position curves for speech errors, collapsed across IS condition, are shown in Figure 3. The characteristic saw-tooth shape is evident for lists of alternating confusability and the overall pattern of errors is very similar to that for the equivalent conditions of Experiment 1.

As for Experiment 1, we classified the speech-errors in terms of the item-types involved (again collapsing over IS conditions). The results are presented in Table 2. This shows that the additional errors on lists containing confusable items stem overwhelmingly from one confusable item's being read in place of another. Of the 462 errors involving C-N substitutions (377 in mixed lists, 85 in pure lists), 176 can be accounted for by confusions between the letters G and J (124 in mixed lists, 52 in pure lists). As noted above, this particular error can also be seen as being promoted by phonemic overlap.

Discussion

The results of Experiment 2 replicated successfully all the important features of those of Experiment 1. Speech errors in the paced reading of six-letter lists mirrored closely the pattern of additional errors attributable to phonological confusability seen in immediate serial recall of the same kinds of list. The saw-tooth pattern in lists of mixed confusability is particularly suggestive of a link between speech errors and phonological confusion errors in ISR. Two aspects of the data-pattern differ across tasks; both concern the effect of serial position. First, in the reading task, it is clear that the first list-position is approximately twice as prone to speech error as neighboring positions; such a pattern is not seen for the effect of confusability in ISR. We have attributed this difference to the demands of the repeated-reading task, in which a gaze shift necessarily precedes all but the first reading of the initial list-item. Second, even disregarding the first list-position, there is no general increase in speech-errors across list positions, as is generally seen in ISR (except for the last position that benefits from a recency advantage). This is naturally explained by noting that the increase in ISR errors across list position is explained in various models (including ours) by a process of ongoing memory decay. In other models this primacy advantage is explained via assumptions regarding output interference on memory. Neither process would be expected to operate in the reading task, however, because participants have no need to remember the list-items, that are plainly visible throughout. The lack of a primacy advantage further supports our assertion that participants in the speech-error elicitation task were not trying to repeat the lists from memory.

It is perhaps for this reason that the manipulation of irrelevant sound had no effect in Experiment 2 (other than perhaps contributing to a slightly higher overall error rate for both speech and noise conditions in this experiment compared with silent conditions in Experiment 1). In our model of the effect (Page & Norris, 2003), irrelevant sound is assumed to affect the order-maintaining primacy gradient that is the first stage of our model. Even if the same order-maintaining stage is engaged as a speech-planning buffer in the speech-error elicitation task (see General Discussion), this stage doesn't seem to be unduly taxed by the reading task. In other words, the fact that the intended order of list items is fixed by the constantly available visual representation of the list (as opposed to decaying or being subject to interference in the ISR task), appears to make the representation of order in the reading task immune to any specific effect of changing state irrelevant sound.

Together, our first two experiments support the hypothesis that the additional errors in lists containing rhyming items can be thought of as onset swaps between the names of the corresponding list items (i.e., akin to contextual speech errors), rather than the more traditionally conceived memory errors in which the relative order of whole list-items is confused. Of course, it might be argued that we are seeing whole-item swaps in both the memory and the speech-error elicitation tasks. There are certainly some errors, such as those on nonconfusable items, that cannot be seen as speech errors resulting from the movement of sublexical units and can only really be seen as whole-item swaps in reading. Nevertheless, the errors on the pure confusable lists are approximately three times as common (over both experiments) as those on pure nonconfusable lists, a fact that would still require explanation even if we conceded that some of the pure confusable reading errors comprised whole-item swaps. The explanation that we are offering, in terms of movement of sublexical segments (onsets) seems the most economical, given the link that it establishes with the speech production literature from which the elicitation task was itself appropriated.

Experiment 3

Our third experiment tackled this issue head on. Here we used factors (other than shared rime or shared vowel) that are known to encourage speech errors. We incorporated these factors into an ISR task (not a speech-error eliciatation task) in which errors resulting from the encouraged speech-error could not be alternatively classified as whole-item exchanges. The two factors we manipulated were phonetic similarity of the onset, and relative frequency of the "Spoonerized" lures. To be explicit, we asked participants to perform ISR of four-word lists. Each four-word list could be thought of as comprising two pairs of two adjacent words, and it was the properties of these pairs that we manipulated. In the condition designed to lure participants into making speech errors (hereafter, the lure condition), the two words in a pair had phonetically similar, but not identical, onset consonants or onset consonant-clusters. For example, if one word began with the phoneme /s/ then the other might begin with a /ʃ/; or if one began with the cluster /dr/ then the other might begin with the cluster /tr/. Both Ellis (1980) and Wilshire (1998, 1999) identified such featural similarity as a factor that encouraged onset-exchange speech errors. The second factor involved consideration of the intended speech error. In the lure condition it was arranged that if participants were lured into making an onset exchange between the members of a word pair, then the two words that would result would both be higher in frequency than either of the two original words. Thus a lure pair might comprise the low-frequency words “bane” and “pelt”, where the intended lures “pain” (or “pane”) and “belt” are of relatively high frequency. Each list was made up of two such pairs and, to control for the unusual nature of some of the low frequency words, the control (nonlure) condition comprised exactly the same words presented in exactly the same list positions, but paired differently in lists such that in nonlure lists no adjacent pairs would form a higher frequency word and, where possible, wouldn’t form a word at all. In this way, any onset exchanges could be unambiguously identified as such, since they would, by hypothesis, result in the intrusion of lexical items (like “pain” and “belt” above) that had not been presented in the stimulus lists.

The use of such lure-pairs in the ISR task also addressed the issue of speech-error frequency. In the introduction we drew attention to the fact that confusability-based order-errors in ISR are much more frequent than are speech errors in normal speech. We suggested that one of the main differences between the representations of serial order in speech production and ISR is whether the order is determined by endogenous or exogenous information. In speech production, order is determined by an endogenous intention that can be maintained throughout recall. In ISR the intended order is determined exogenously by the presentation list and this order information is subject to fairly rapid decay and/or interference. Because ISR draws on a degraded speech plan, in the current ISR experiment we should see a frequency of induced *speech errors* that is more like that which we hypothesize in the PSE than that which we might expect to see in everyday speech. We would not, however, expect to see as large an effect as in the PSE: with up to five rhyming lures (in a six-item list) including some with phonetically similar onsets (e.g., B and P, T and D), the PSE involves a particularly strong manipulation of the factors that we believe are at work.

Method

Participants.

The participants were 18 members of the MRC Cognition and Brain Sciences Unit's subject panel. There were 13 females and 5 males with a mean age of 18.6 years. Each was paid a small fee (£5) for taking part in the experiment.

Materials.

Participants were presented with 64 lists, each comprising four words for immediate serial recall. Pairs of single syllable words were selected such that they shared some featural similarity in their onset consonants or consonant clusters. Words were chosen with regard to frequency (as recorded by Kucera & Francis, 1967), such that, as far as possible, all stimulus words had a frequency of fewer than 15 occurrences per million words. This upper frequency limit was breached in only two cases, both of which resulted from failure to consider homophones. Thus, our chosen word “tor” is of zero frequency, but its homophones “tore” and “tour” are much more frequent (15 and 43 occurrences per million, respectively). Likewise, “par” (13 occurrences) met the frequency criterion, but its homophone “pa” (32 occurrences) did not. Excluding these two outliers, the mean frequency of the remaining 62 stimulus words was 2.5 occurrences per million (range 0–14, s.d.= 3.1, median = 1) where, in the case of homophones, the sum of the homophone frequencies was used. As noted above, in the lure condition, the words were paired so that an exchange of onset (cluster) would result in two words of relatively high frequency. The mean frequency of these lure words was 150 occurrences per million (range 13–1600, s.d. = 267, median = 75), with only one lure ("rug") having a frequency of fewer than 15 occurrences per million.

Lists in the lure condition, therefore, comprised two consecutive word pairs constructed as above, such that the members of a pair shared featural similarity in their onsets and would produce two higher frequency words if Spoonerized. No attempt was made to prevent adjacent-item, cross-pair lures, that is words that could be formed by Spoonerizing items 2 and 3 in a list (although featural onset similarity in these cases was rare). There were 22 words that could be formed by Spoonerizing list items 2 and 3, and these lures had a mean frequency of 142 occurrences per million (median = 15).

In the nonlure lists, exactly the same words were used, and in exactly the same within-list positions as in the lure condition, the difference being that words were paired differently (in positions 1 and 2, and in positions 3 and 4) such that they would not,as far as was possible, share onset features and they would not Spoonerize to make words of relatively high frequency. It was not possible completely to meet these conditions: in 10 cases, the onset consonants of two paired words shared all but one major class feature - in none of these cases, however, would a pair of words have resulted from the onset exchange; in 20 cases, a word could result from an onset exchange between members of a pair of words in the nonlure condition. The resulting words had a mean frequency of 39 (s.d. = 78, median = 4) occurrences per million, considerably lower than the equivalent words in the lure condition. In addition, there were 8 different words that could be formed by an onset exchange between the adjacent items in list positions 2 and 3 of the nonlure condition; these had a mean frequency of 20 occurrences per million (s.d. = 37, median = 3), again a smaller number of words, and of considerably lower frequency, than those in the equivalent positions of the lure condition.

All the words for the experiment, together with the word “Recall”, were digitally recorded in mono at a 44.1kHz sample rate. They were spoken by a single male speaker, at a regular rate of 1 per 750ms in single take, with every effort made to ensure a monotone throughout. The recording was then spliced into individual 750ms sections, each containing a word, the location of whose perceptual center (the perceived moment of occurrence) within the extracted portion was equated as closely as possible across words. This provision was necessary to ensure that the words sound naturally paced in any reordering, and its success was confirmed by listening to the lists as finally presented.

Lists in the two conditions, lure and nonlure, were randomly intermixed for presentation to participants. Each word was used four times, occurring in two lure lists and two nonlure lists. The complete set of 64 lists that resulted is shown in Appendix A.

Procedure.

Lists were presented by computer, in the same order to all participants. Participants commenced each trial by pressing the spacebar. One second later, the four words were presented at a presentation rate of one word per second, followed one second later by the word “Recall”. The verbal cue to recall was inserted so as to reduce the size of any modality effect (an advantage for auditory over visual presentation), that might otherwise have moved performance too close to ceiling levels of performance, particularly for the final list-item. Participants then spoke their response, attempting to recall each word in order, and saying “blank” if they wished to omit a word in a given response position. Participants’ responses were digitally recorded in the same manner as in Experiments 1 and 2, and their responses were marked subsequently, as described below.

Results

The marking of participants’ responses proceeded by classifying each of the responses as one of the following: a correct response, comprising a word in its correct position; a transposition error, comprising a word from the stimulus list recalled in the wrong position; an omission (“blank”); an adjacent spoonerism, comprising an intruded word made up from an onset and rime taken from adjacent items in the stimulus list; a nonadjacent spoonerism, comprising an onset and a rime taken from anywhere in the list other than from adjacent items; an intrusion, comprising a new word or nonword that fitted none of the above definitions. Our predictions were two-fold: that overall performance would be worse for the lure condition than for the nonlure condition; and that this difference would largely be seen in the number of adjacent Spoonerisms that participants were lured into making in the lure condition.

Scores in each category were subjected to separate two (condition: within) by four (serial position: within) repeated-measures analyses of variance. The effect of serial position was reliable in all analyses except that for nonadjacent Spoonerisms, but this factor never interacted reliably with condition. In presenting the results in Table 3, therefore, we have collapsed across serial positions to give a proportionate score in the two conditions for each class of response. These are accompanied in each case by the F-value for the main effect of lure condition, and its corresponding p-value. As can be seen, the predictions were borne out. Participants performed worse in the lure condition than in the nonlure condition and the only error category that distinguished between conditions was that relating to adjacent Spoonerisms. It is important that both effects were found. In the nonlure condition, most adjacent Spoonerisms would result in nonwords. It is crucial, therefore, that the increased number of Spoonerisms in the lure condition, was not offset by a compensating increase in intrusions and/or omissions in the nonlure condition. This implies that the factors that we have manipulated, namely overlap in onset features and lure frequency, were instrumental in promoting the occurrence of errors, rather than simply affecting the way in which a given number of errors was classified.

Discussion

The results of this experiment are relatively straightforward. By manipulating factors that are known to promote speech errors, we have been able to promote the occurrence of errors in an immediate serial recall task. The manipulation produced a 4.6% difference in performance in a situation in which mean performance was over 82%. In the lure condition, onset movements, that could not in these circumstances be mistaken for whole-item movements, occurred on 8% of responses.

Taken together, these results imply that the immediate serial recall task is particularly prone to onset-movement errors, that would in other circumstances be classified as speech errors. The results of Experiment 3 demonstrate that if we manipulate factors that have been shown elsewhere to produce speech errors, then we will see such errors emerging relatively strongly in the context of immediate serial recall. The traditional use of rhyming items, in previous demonstrations of the PSE, is a particularly strong manipulation of factors liable to produce speech errors: onset movements are promoted by strong context similarity (shared rime across a span of items), often by featural similarity between onsets (e.g. B/P and T/D), and by the fact that the results of any onset exchange will not only be relatively frequent items (as in Experiment 3) but will be the stimulus items themselves, just reordered. In our view, therefore, it is reasonable to expect at least as many onset exchanges in ISR of a list of rhyming items as we saw for the nonrhyming items of Experiment 3. If this is accepted, then it follows that onset exchanges, rather than whole-item exchanges, will make up a significant proportion, if not all, of the additional errors seen in the ISR of lists of rhyming items.

General Discussion

The data reported here are consistent with the hypothesis that the additional serial order errors found in ISR of phonologically similar (often rhyming) items, share a locus with a class of speech error in which sublexical units move between words. It is, of course, possible that the similarity in error patterns is coincidental, and that errors in the ISR of rhyming items might only resemble sublexical errors in everyday speech, rather than being due to a common mechanism. However, adopting such a perspective has two drawbacks. First, it is less parsimonious to postulate two separate mechanisms having exactly the same properties, rather than a single store. Second, a two-mechanism account would still leave us without any explanation for why errors in ISR behave in a way that can only be explained in terms of two-stage models when, as noted above, the use of two stages confers no advantage in ISR. As we argued in the Introduction, this otherwise puzzling character of ISR behaviour makes sense under the assumption that the ISR task recruits the existing speech production machinery. In what follows, we work through the theoretical consequences of the assumption that ISR and speech production share such machinery.

The most far reaching consequence can be summarized in a question: If the output stage in ISR is assumed to be the same as that used in the later stages of speech production, then why not further assume that the first stage of that model is also shared with an earlier stage of speech production? Note that this is not a necessary extension: it might just be that the memory system underlying verbal ISR feeds into the speech production system at a point at which the phonology of to-be-produced words is being constructed. Nonetheless, the idea is an enticing one, and is certainly what Ellis (1980) intended when he identified the phonological store with "a response buffer, whose normal function is to allow efficient programming of speech production by holding preplanned stretches of impending speech in the interval between an utterance or part of an utterance being planned and its being overtly articulated". Put simply, the idea is that, in the context of speech reproduction, the phonological store *is* the high-level lexical plan that drives the utterance of an ordered string of verbal material.

This assumption turns out to be compatible both with models of speech production and our own model of ISR (Page & Norris, 1998). Taking the models of Dell (1986, 1988) and Levelt et al. (1999) as representative, it is clear that though speech production models have been primarily directed at simulating the production of single words or two-word phrases, their authors have always kept in mind the ultimate requirement to generate multiword utterances. Dell, et al. (1997) were particularly clear in relating speech production to theories of serial memory in general and, as noted earlier, drew attention to the fact that the particular error patterns seen in speech production require that future events (e.g., upcoming words) are partially activated during the planning of the current word. Houghton (1990) made many of the same points. The primacy gradient of activation that lies at the heart of our model of ISR is a good candidate for an activation-based mechanism that instantiates exactly this property, that upcoming words are activated to an extent that depends on their imminence. Looked at in this way, the primacy gradient constitutes a specification for the order in which the elements of a planned utterance should be generated. This is precisely what is required to elaborate models of speech production that have hitherto concentrated on single-word production.

With regard to the ISR paradigm, the idea is as follows. Faced with, say, a sequence of visually presented items for immediate serial recall, the participant takes advantage of the early (word ordering) stages of a speech production system that is exquisitely designed to produce words strictly in the intended order. The participant recodes the visual stimuli into a speech-based form, and maybe even covertly repeats partial sequences (rehearsal). This converts the visual sequence into a planned utterance encompassing the relevant words in the correct order. Unlike normal utterances in which the participant engages, there is no constraining semantic or syntactic content associated with the utterance and there is, therefore, nothing to prevent decay of (or interference with) the plan. This leads to poorer order memory and, as described above, more sublexical errors than would be seen in more typical, endogenously driven speech. Of course, for visual presentation in the presence of concurrent articulation, recoding will be impossible and there will thus be no *speech* to reproduce. These are the circumstances in which the PSE is abolished, as would be expected.

For an auditory stimulus, recoding is unnecessary. According to the working memory model auditory material can be directly encoded into the phonological store. The equivalent statement, couched in terms of a speech production system (acting, in this case, as a speech *reproduction* system), is that we possess a system that relatively automatically establishes an utterance plan that corresponds to (a latter portion of) the sequence of words that we have just heard. Storage in the phonological store would, accordingly, be conceived as the set of processes that gives rise to the ordered representation of a recently heard (or recoded) sequence, such that that representation is able to drive directly the production of an utterance comprising the same sequence.

Having outlined the relationship between our modified perspective of the phonological loop and its classical predecessor, we are in a position very briefly to address two questions that have recently been raised by Jones et al. (2004): Is the phonological store phonological? And is it a store? Dealing with the latter first, to the extent that ISR demands that speech-output processes reproduce a recently heard (or recoded) stimulus, then they must act on a representation of the stimulus that preserves (i.e., stores) information relating to items and their serial order. The phonological store is indeed a store, therefore, albeit a rather labile one. Turning to the first question, to the extent that the machinery underlying ISR of verbal materials is as intimately speech-based as our analysis suggests, then it is, broadly speaking, “phonological”, depending on how one defines the term. Of course, provision of a precise quantitative model rather mitigates extended debate over terminology.

Finally, we come to an important issue regarding neuropsychological data. At first sight our hypothesis, regarding the identity of the phonological store and a lexical-level utterance plan, faces a severe problem here. This problem stems from the observation that so-called short-term memory patients (Basso, Spinnler, Vallar, & Zanobio, 1982; de Renzi & Nichelli, 1975; Trojano & Grossi, 1995; Warrington & Shallice, 1969), whose ability to perform auditory immediate serial recall is severely compromised, do not typically have any difficulty in producing everyday speech. If their lexical ordering mechanisms are functional in the execution of everyday speech, and we have identified those mechanisms with the phonological store, then why are these patients so impaired in their auditory sequence span?

Our explanation involves once again drawing attention to the difference between endogenous and exogenous speech-output. The auditory serial recall task involves repeating a recently heard sequence. In the discussion above, we have hypothesized that the direct access to the phonological loop, classically assumed for auditory material, depends on processes that are able automatically to convert an incoming auditory sequence into the primacy gradient corresponding to the ordered plan for production of the equivalent sequence. One such mechanism is described in our earlier papers (Page & Norris, 1998a, 1998b), though the precise detail is less important than the realization that such processes are a necessary component of the story. Importantly, however, such conversion processes are not necessary in the production of endogenous speech.

The planning of endogenous speech, that is to say the loading up in the relevant sequential plan based on an intended message, does not involve the conversion of an incoming stimulus. Of course, such planning does involve other semantic and syntactic mechanisms, but these are separate from those involved in the conversion of incoming lists. Given this distinction, we are able to solve the apparent problem raised by the STM patients, by assuming that what is lacking in these patients is the ability to convert an incoming auditory sequence into the primacy gradient appropriate for its repetition. What is not lacking is the substrate for instating primacy gradients at all, and it is this that they are able to use in everyday speech production.

Conclusion

This paper builds on our previous work (Page & Norris, 1998a, 1998b) in explicitly bringing speech-error data together with those relating to the phonological similarity effect in immediate serial recall. No current model of immediate serial recall (other than our extended model, whose predictions are being tested here) has maintained that the PSE is a result of sublexical movements. The additional order errors that result from phonological similarity have previously been ascribed to whole-item movements, omissions or intrusions. With regard to theories of speech production , we believe that the data presented above, particularly those from Experiment 3, are indicative of the particular susceptibility of immediate serial recalls to certain classes of sublexical speech error. We see far more sublexical movements in the recall of a four-word list than we would expect in, say, a single reading of the same. It has often been observed informally that participants will blend two words in an ISR task (e.g., “yacht” and “goose” might be blended to produce “got”), but the observation has not hitherto been so closely controlled as here, nor has it been so closely related to the PSE itself. The fact that ISR is so susceptible to sublexical movements might well permit new experimental manipulations of factors thought to be involved in the genesis of such speech errors, by providing a paradigm in which higher rates of speech error can be expected.

The speech-error elicitation task has not, to our knowledge, been applied to the sorts of letter-list stimuli that one typically uses in studying the PSE in immediate serial recall. In particular, we know of no study that has used either six consecutive rhyming items, or lists of alternating rhyme, in a speech-error elicitation task. The fact that we have used the same lists in both tasks here, and shown a distinct similarity in the error patterns obtained is, therefore, a novel contribution. Traditional tongue-twisters often rely on featural similarity between syllable onsets for their effect. Our paper has focussed instead on the difficulties associated with uttering a series of rhyming items. It has been known for many years that such sequences are difficult to recall: we have now explicitly attributed this fact to a demonstrable difficulty in production.

Finally, we believe that the discussion above, regarding patient populations and the topology of the language/memory system, is both original and valuable. It is often assumed that the fact that so called “short-term memory patients” are perfectly able to speak in everyday life, is sufficient evidence against our hypothesis of an identity between an abstract (lexical-level) utterance plan and the phonological loop component of working memory as manifested in the recall of a list of words. We have argued that this assumption is incorrect.

In summary, we have presented evidence that corroborates the hypothesis that errors underlying the phonological similarity effect in immediate serial recall share a locus with sublexical errors in everyday speech. This evidence builds on work by Morton (1970) and Ellis (1980), as well as on our own earlier work (Page & Norris, 1998a, 1998b), while maintaining contact with the classical working memory perspective (e.g., Baddeley, 1986) by which our work has been, and continues to be, strongly influenced.

References

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. Quarterly Journal of Experimental Psychology, 18, 362-365.

Baddeley, A. D. (1968). How does acoustic similarity influence short-term memory? Quarterly Journal of Experimental Psychology, 20, 249-264.

Baddeley, A. D. (1986). Working memory. New York: Oxford University Press.

Baddeley, A. D. (2004, July). Paper presented at the Working Memory Discussion Meeting, Caer Llan.

Baddeley, A. D., & Hitch, J., Graham. (1974). Working memory. In G. Bower (Ed.), The psychology of learning and motivation, vol. 8 (p. 47-90). London: Academic Press.

Baddeley, A. D., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. Quarterly Journal of Experimental Psychology, 36A, 233-252.

Basso, A., Spinnler, H., Vallar, G., & Zanobio, M. E. (1982). Left hemisphere damage and selective impairment of auditory verbal short-term memory: A case study. Neuropsychologia, 20, 263-274.

Burgess, N., & Hitch, G. J. (1992). Toward a network model of the articulatory loop. Journal of Memory and Language, 31, 429-460.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. Psychological Review, 106, 551-581.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. Journal of Verbal Learning and Verbal Behavior, 15, 17-31.

Conrad, R. (1964). Acoustic confusions in immediate memory. British Journal of Psychology, 55, 75-84.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. British Journal of Psychology, 55, 429-432.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological Review, 93, 283-321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. Journal of Memory and Language, 27, 124-142.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. Psychological Review, 104, 123-147.

Ellis, A. W. (1980). Errors in speech and short-term memory: The effects of phonemic similarity and syllable position. Journal of Verbal Learning and Verbal Behavior, 19, 624-634.

Fallon, A. B., Groves, K., & Tehan, G. (1999). Phonological similarity and trace degradation in the serial recall task : When CAT helps RAT, but not MAN. International Journal of Psychology, 34, 301-307.

Farrell, S., & Lewandowsky, S. (2003). Dissimilar items benefit from phonological similarity in serial recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 838-849.

Hartley, T., & Houghton, G. (1996). A linguistically constrained model of short-term memory for nonwords. Journal of Memory and Language, 35, 1-31.

Henson, R. N. A. (1998). Short-term memory for serial order: The Start-End model. Cognitive Psychology, 36, 73-137.

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 49A, 80-115.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), Current research in natural language generation (p. 287-319). London: Academic Press.

Jones, D. M., Macken, W. J., & Nicholls, A. P. (2004). The phonological store of working memory: Is it phonological and is it a store? Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 656-674.

Jones, D. M., Madden, C., & Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: The role of changing state. Quarterly Journal of Experimental Psychology, 44A, 645-669.

Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence: Brown University Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-75.

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, *8*, 323-350.

Morton, J. (1964). A preliminary functional model for language behaviour. *International Audiology*, *3*, 215-225.

Morton, J. (1968). Grammar and computation in language behaviour. In J. Catford (Ed.), *Studies in language and language behaviour.* Anne Arbor: Univ. of Michigan Press.

Morton, J. (1970). A functional model for memory. In D. Norman (Ed.), *Models of human memory.* New York: Academic Press.

Nairne, J. S., & Kelley, M. R. (1999). Reversing the phonological similarity effect. *Memory and Cognition*, *27*, 45-53.

Nimmo, L. M., & Roodenrys, S. (2004). Investigating the phonological similarity effect: Syllable structure and the position of common phonemes. *Journal of Memory and Language*, *50*, 245-258.

Nooteboom, S. G. (1967). Some regularities in phonemic speech errors. *Institute for Perception Research, Eindhoven, Annual progress Report*, *2*, 65-70.

Norris, D., Page, M. P. A., & Baddeley, A. D. (1994, July). *Serial recall: it's all in the representations.* Paper presented at the International Conference on Working Memory, Cambridge, England.

Page, M., & Henson, R. (2001). Computational models of short-term memory: Modelling serial recall of verbal material. In J. Andrade (Ed.), *Working memory in perspective* (p. 177-198). New York: Psychology Press.

Page, M., & Norris, D. (1998b). Modeling immediate serial recall with a localist implementation of the primacy model. In A. M. Jacobs & J. Grainger (Eds.), *Localist connectionist approaches to human cognition.* (p. 227-255). Mahwah, NJ: Lawrence Erlbaum Associates.

Page, M. P. A., & Norris, D. (1998a). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761-781.

Page, M. P. A., & Norris, D. G. (2003). The irrelevant sound effect: What needs modelling, and a tentative model. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *56A*, 1289-1300.

Renzi, E. de, & Nichelli, P. (1975). Verbal and non-verbal short-term memory impairment following hemispheric damage. *Cortex*, *11*, 341-354.

Stemberger, J. P. (1990). Wordshape errors in language production. *Cognition*, *35*, 123-157.

Trojano, L., & Grossi, D. (1995). Phonological and lexical coding in verbal short-term memory and learning. *Brain and Language*, *51*, 336-354.

Warrington, E. K., & Shallice, T. (1969). The selective impairment of auditory verbal short-term memory. *Brain: A Journal of Neurology*, *92*, 885-896.

Wickelgren, W. A. (1965). Short-term memory for phonemically similar lists. *American Journal of Psychology*, *78*, 567-574.

Wilshire, C. E. (1998). Serial order in phonological encoding: An exploration of the 'word onset effect' using laboratory-induced errors. Cognition, 68, 143-166.

Wilshire, C. E. (1999). The tongue twister paradigm as a technique for studying phonological encoding. Language and Speech, 42, 57-82.

Appendix A

These were the lists used in Experiment 3, classified by list type.

POX TOME SHINE RUCK (nonlure)

RACK GUT LICE COT (nonlure)

SOT SHOAL TOR DENT (lure)

POX BOAST SHINE SIP (lure)

RACK LOOM LICE ROOK (lure)

NODE GAUZE TEAL SHINE (nonlure)

VAULT FEIGN MOAT NARK (lure)

PAR BANE TOME DELL (lure)

GOAT LEACH SHUN DENT (nonlure)

CRANE MOAT LEER SOP (nonlure)

BANE PELT PEACH BARK (lure)

COT GAUZE RUCK LEER (lure)

TOUT DIME NARK MAIM (lure)

CRANE GRAFT RYE LUG (lure)

SHALE SIFT JEST CHUMP (lure)

LOOT RAND GRAFT CRANE (lure)

MEAD NAY BOAST POX (lure)

LOOT PELT GRAFT MAIM (nonlure)

DENT PEACH MOAT SIP (nonlure)

GRIME JEST SHOAL TOG (nonlure)

COD GOAT DIME TOUT (lure)

DRIP TRESS GAUZE COT (lure)

NAY SHOAL LUG DELL (nonlure)

DALE CURL SIFT GRIME (nonlure)

NAY MOAT LEER RUCK (lure)

BANE TEAL GASH DRIP (nonlure)

COD BANE DIME RACK (nonlure)

SHALE LOOM JEST DEEM (nonlure)

RAND LOOT PELT BILE (lure)

DELL RAND TOG CRANE (nonlure)

GUT CURL CRAVE GRIME (lure)

FEIGN VAULT TRESS DRIP (lure)

CHUMP JEST LEACH ROT (lure)

DELL TOME TOG DEEM (lure)

ROT TRESS DEEM NODE (nonlure)

GOAT COD TEAL DALE (lure)

SOT GRAFT TOR CANE (nonlure)

TOUT FEIGN NARK CHUMP (nonlure)

BILE PAR GASH CANE (lure)

ROT LEACH LUG RYE (lure)

GUT LOOT CRAVE BARK (nonlure)

NODE MEAD SIFT SHINE (lure)

ROOK VAULT RYE GASH (nonlure)

CURL NAY SHIN LEER (nonlure)

VAULT TOR PELT LUG (nonlure)

SOCK GOAT SIP BILE (nonlure)

DALE TEAL SHUN SOP (lure)

FEIGN BOAST TRESS ROOK (nonlure)

MEAD SHUN CANE TOUT (nonlure)

CHUMP COD LEACH SHALE (nonlure)

CURL GUT SHIN SOCK (lure)

BARK PEACH LOOM RACK (lure)

RAND PAR MAIM ROT (nonlure)

GRIME CRAVE SHOAL SOT (lure)

DENT TOR MAIM NODE (lure)

PAR LICE TOME SOT (nonlure)

BARK CRAVE LOOM POX (nonlure)

SOP SHUN CANE GASH (lure)

ROOK LICE DEEM TOG (lure)

COT SHIN RUCK NARK (nonlure)

BILE SIFT PEACH SOCK (nonlure)

SOCK SHIN SIP SHALE (lure)

DRIP MEAD GAUZE RYE (nonlure)

SOP DIME BOAST DALE (nonlure)

Author Notes

We would like to thank Dr Mike Greenhough, Jane Hall, and Jon Williams for their assistance in running the experiments reported here. All correspondence and requests for reprints should be sent to Dr Mike Page (m.2.page@herts.ac.uk), Dept. of Psychology, University of Hertfordshire, Hatfield, AL10 9AB, UK.

Table 1: Classification of item exchanges, scaled by number of opportunities

List-type	Exchange type		
	C with C	C with N	N with N
Alternating-CN	59 (177/3)	11 (99/9)	6 (19/3)
Alternating-NC	51 (153/3)	12 (109/9)	2 (6/3)
Pure C	41 (618/15)	— (46)	— (0)
Pure N	— (0)	— (13)	7 (111/15)

Table 2: Classification of item exchanges, scaled by number of opportunities

List-type	Exchange type		
	C with C	C with N	N with N
Alternating-CN	69 (206/3)	17 (156/9)	21 (63/3)
Alternating-NC	101 (304/3)	25 (221/9)	6 (17/3)
Pure C	58 (871/15)	— (16)	— (0)
Pure N	— (0)	— (69)	14 (207/15)

Table 3: Mean proportions error (s.e.) for the conditions of Experiment 3

Error type	Condition		F(1,17)	p
	Nonlure	Lure		
Overall	0.156 (0.027)	0.202 (0.025)	6.81	.02
Transpositions	0.022 (0.005)	0.017 (0.005)	0.52	.48
Omissions	0.043 (0.015)	0.047 (0.015)	0.29	.60
Adjacent Spoonerisms	0.018 (0.004)	0.076 (0.010)	29.9	<.001
Nonadjacent Spoonerisms	0.007 (0.001)	0.010 (0.002)	2.13	.16
Intrusions	0.066 (0.010)	0.053 (0.009)	2.68	.14

Figure Captions

Figure 1. Serial position curve for the ISR task in Exp. 1 (errorbars represent plus/minus one standard error of the mean).

Figure 2. Serial position curve for the speech error elicitation task in Exp. 1.

Figure 3. Serial position curve for the speech error elicitation task in Exp. 2.

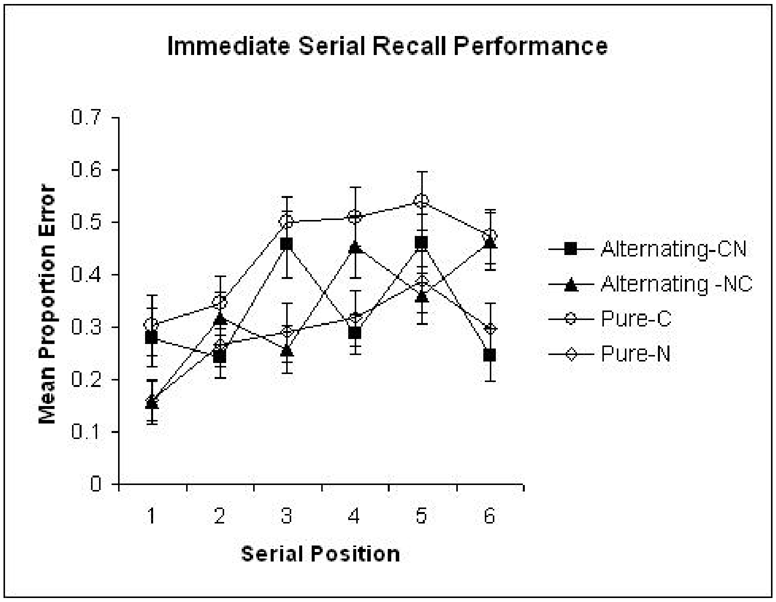
Immediate Serial Recall Performance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean Proportion Error
1
2
3
4
5
6
Serial Position
Alternating-CN
Alternating -NC
Pure-C
Pure-N

Figure 1:

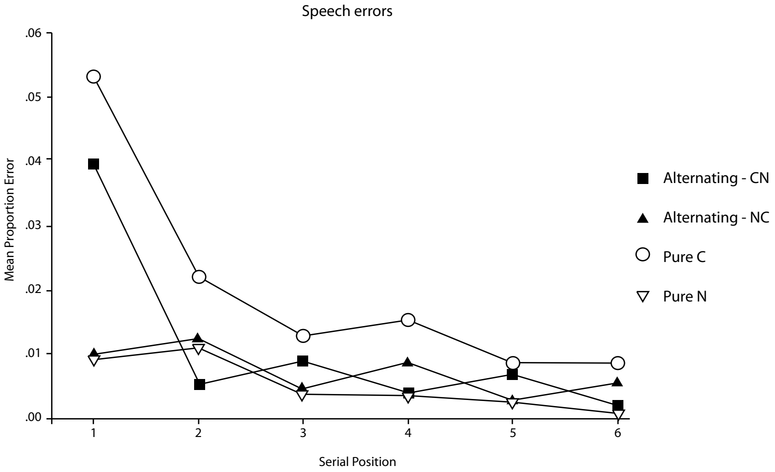
Speech errors
Mean Proportion Error
.06
.05
.04
.03
.02
.01
.00
1
2
3
4
5
6
Serial Position
■ Alternating - CN
▲ Alternating - NC
○ Pure C
▽ Pure N

Figure 2:

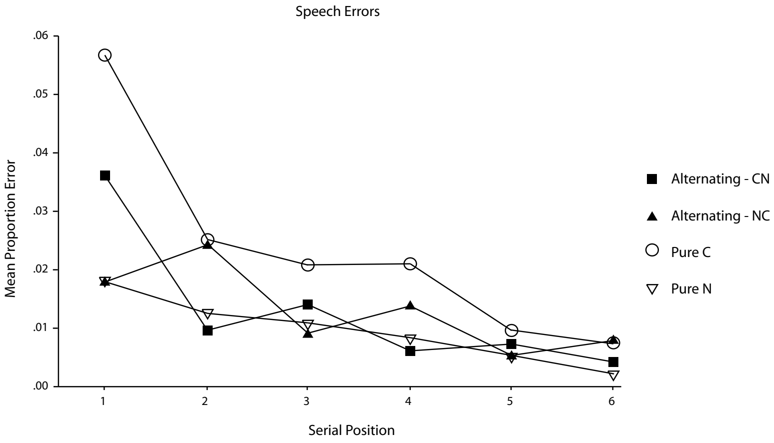
Speech Errors
Mean Proportion Error
.06
.05
.04
.03
.02
.01
.00
1
2
3
4
5
6
Serial Position
■ Alternating - CN
▲ Alternating - NC
○ Pure C
▽ Pure N

Figure 3: