---

**Book Section:**
Hartley, Tom (2022) Computational Models of Working Memory for Language. In: Schwieter, John W. and Wen, Zhisheng, (eds.) Cambridge Handbook of Working Memory and Language. Cambridge University Press.

https://doi.org/10.1017/9781108955638

---

# Computational Models of Working Memory for Language

Graham J. Hitch
University of York

Mark J. Hurlstone
Lancaster University

Tom Hartley
University of York

We start with a brief review of evidence that verbal working memory (WM) involves a limited capacity phonological loop capable of retaining verbal sequences for a few seconds in immediate serial recall, vocabulary acquisition, speech production and language comprehension. The challenge of explaining how such a system handles information about serial order is discussed in the context of computational models of the immediate recall of unstructured sequences of words, letters or digits, an extensively studied laboratory task for which there are many benchmark findings. Evaluating computational models against these benchmarks suggests a serial ordering mechanism in which items are simultaneously active before being selected for sequential output by a process of competitive queuing (CQ). Further evidence shows how this process may operate in the context of sequences that conform to various kinds of linguistic constraint. We conclude by suggesting that CQ is a promising theoretical mechanism for connecting and potentially unifying theories of WM and language processing more generally despite major differences in their scope and level of abstraction.

## Introduction

We begin by setting out briefly our view of working memory (WM) as a system, and how it stores verbal inputs. As a system, WM refers to the limited capacity cognitive resources we draw on in everyday activities such as mental arithmetic (Hitch, 1978) and following instructions (Yang, Allen, Gathercole, 2015) in which temporary information has to be both kept in mind and operated on. Theoretical accounts agree that WM involves an attentional resource that interacts with a set of transient memory representations in order to achieve some goal (Logie, Camos, Cowan, 2021). However, beyond this broad consensus there are many areas of disagreement and emphasis. For example, most approaches view WM as a general-purpose system while some see it as specialised for language (e.g., Schwering MacDonald, 2020). However, this is probably best thought of as a difference in emphasis given evidence that verbal and nonverbal tasks draw at least in part on a common pool of limited resources (Kane, Hambrick, Tuholski, Wilhelm, Payne, Engle, 2004; Morey, 2018). Another point of difference is between approaches in which transient information is assumed to correspond to currently activated information in long-term memory (LTM) (Cowan, 1995; Oberauer, 2002) and those in which it reflects the operation of short-term buffer stores (Baddey, 2000; Baddeley Hitch, 1974; Logie, 1995). We will argue that this difference is also best thought of as one of emphasis and that WM involves both activated LTM and the temporary storage of novel information. The crucial questions concern how these interact and combine and we will go on to show that attempts to develop computational models allow us to test specific proposals. We begin by discussing the evidence for a component of WM capable of storing novel verbal input in phonological form over brief intervals as we regard this capacity as central to its role in language processing.

## Transient Storage in Verbal WM

The idea that recent perceptual inputs are briefly represented in a limited capacity buffer store was first developed by Broadbent (1958) and was an important feature of the Modal Model of memory, in which short-term memory (STM) was regarded as the gateway to LTM (Atkinson Shiffrin, 1968). At that time several strands of evidence were interpreted as supporting the idea of separate short-term and long-term memory stores. These included the limited span of immediate recall (Miller, 1956) coupled with the absence of a measurable limit on LTM; instances of selective neuropsychological impairment to STM in combination with preserved LTM (Shallice Warrington, 1970) and the converse (Baddeley Warrington, 1970); separate short-term and long-term components in free recall (Glanzer Cunitz, 1966); and use of acoustic coding in STM and semantic coding in LTM (Baddeley, 1966a, 1966b). However, despite initial acceptance, certain assumptions in the Modal Model failed to stand up to scrutiny. For example, Shallice and Warrington's (1970) STM patient showed normal long-term learning, challenging the idea of STM as the gateway to LTM. There was also evidence that identifying STM with acoustic coding and

LTM with semantic coding was overly simplistic (Shulman, 1972; Nelson Rothbart, 1972). Unsurprisingly, the concept of STM became unfashionable. Research interest switched to learning and LTM (Craik Lockhart, 1972) and there were influential statements of STM's demise (Crowder, 1982).

To some, however, it was premature to abandon interest in STM given that Atkinson and Shiffrin (1968) saw it as serving the function of WM. Evidence that neuropsychological impairment of STM is not associated with general intellectual impairment (Shallice Warrington, 1970) casts doubt on this suggestion. Nevertheless, Baddeley and Hitch (1974) considered it worth exploring further. They carried out a series of dual-task experiments to examine people's ability to perform a range of cognitive tasks when STM is loaded with irrelevant verbal information. The cognitive tasks involved verbal reasoning, comprehending prose, and long-term learning in verbal free recall. The outcome was a consistent pattern across all these tasks in that small loads could be maintained with little interference and even when STM was loaded to capacity the disruption was mild and far from catastrophic. This suggested that tasks loading STM tap into only one aspect of WM. To capture this Baddeley and Hitch (1974) proposed a multi-component model of WM in which a limited capacity attentional resource directs control processes over two modality-specific buffer stores, one corresponding broadly to verbal STM, the other to visuo-spatial STM. The underlying philosophy was to set out a broad framework capable of generating further questions that, if fruitful, would lead to progressive refinements to the model. Fortunately, this has proved to be the case in that the original framework has undergone a series of revisions in the light of further research (Baddeley, 1986; Baddeley, 2000; Baddeley, Allen, Hitch, 2011) and is still widely used, despite having to compete with an ever increasing number of alternative models (Logie, Camos, Cowan, 2021).

## Phonological Loop

In the multi-component model of WM, the verbal buffer was initially regarded as a temporary store for speech-coded information. Information in this store was assumed to fade rapidly unless refreshed by vocal rehearsal that could be either overt or covert. The principal evidence for speech-coding was the observation that immediate serial recall is poorer when items are phonologically similar even when presented visually (Conrad, 1964). Evidence for the role of covert rehearsal came from the observation that suppressing articulation by repeating an irrelevant word impairs immediate serial recall and at the same time removes the phonemic similarity effect (Murray, 1968). Another line of evidence came from the word length effect, the observation that immediate serial recall declines as the time needed to articulate items is increased, an effect that is also removed by articulatory suppression (Baddeley, Thomson, Buchanan, 1975). Taken together these observations could be explained by assuming information in the phonological loop decays in approximately 2 s unless refreshed by subvocal rehearsal. Subsequent findings have challenged the assumption that rapid forgetting is due to decay (Jalbert, Neath, Bireta, Suprenant, 2011; Lewandowsky, Oberauer, Brown, 2009; Service, 1998). However, this does not undermine the idea that rapid forgetting happens, only the mechanism through which it occurs, and the basic concept remains influential.

The operation of the phonological loop had to be spelled out in more detail to address effects of the modality used to present items for immediate recall. Thus, with visual presentation, articulatory suppression abolishes both the phonemic similarity and word length effects whereas with spoken presentation, suppression abolishes only the word length effect, leaving that of phonemic similarity intact (Baddeley, Lewis, Vallar, 1984). This is consistent with spoken items accessing the loop automatically whereas visual items have first to undergo phonological coding, which depends on subvocalisation. These effects of presentation modality emphasise a view of the phonological loop as a system specialised for speech input-output.

Attempts to investigate recoding in more detail suggest the need to distinguish between different types of phonological code. Thus, whereas articulatory suppression disrupts the ability to decide whether printed words such as BLAME and FLAME rhyme it has little effect on the ability to decide whether words such as AIL and ALE are homophones or whether a nonword such as PALLIS sounds like a real word (Besner, Davies, Daniels, 1981; see also Baddeley, Eldridge, Lewis, 1981). One interpretation of these differences is that a process of orthographic to phonological recoding feeds into, but is distinct from, the process of subvocal rehearsal. This "inner ear" is capable of supporting holistic judgements of homophony whereas the "inner voice" of subvocalisation is required to allow the segmentation operations involved in making rhyme judgements (Baddeley et al., 1981; see also Vallar Papagno, 2002). Inner speech has been shown to activate neurophysiological motor-to-auditory mappings which together with auditory-to-motor mappings may form the neural substrate of the phonological loop (see e.g., Ylinen et al., 2015).

## Activated LTM

Having briefly described the evidence for the phonological loop as a buffer for speech input-output it is instructive to consider the alternative view that these phenomena reflect the transient activation of representations in LTM. The principal drive for this type of account comes from evidence for substantial effects of prior long-term learning in STM tasks. To give some examples, immediate memory is better for words than nonwords (Hulme, Maugham, Brown, 1991), for words with higher frequencies of occurrence in the lan-

guage (Gregg, Freedman, Smith, 1989), and for nonwords with higher phonotactic frequencies (Gathercole, Frankish, Pickering, Peaker, 1999). It is possible to explain some of these effects in terms of redintegration processes at retrieval that use information in LTM to "clean up" degraded phonological representations (Hulme et al., 1991; Schweickert, 1993). However, more generally this seems an implausible way to account for phenomena such as the immediate recall of meaningful sentences, where word span is typically measured in double figures (Brener, 1940). A more appealing hypothesis is that performance in verbal STM tasks reflects the interactive activation of semantic, syntactic, lexical and phonological information according to the combinatorial statistics of language knowledge (Martin Saffran, 1997). Proponents of integrative accounts of WM and language push this approach to the limit by assuming no role for a phonological or indeed any other type of short-term buffer (see e.g. McDonald, 2016; Schwering MacDonald, 2020). However, Norris (2017) pointed out the necessity of assuming some form of transient storage that goes beyond the mere reactivation of existing representations in LTM. This is needed to explain our ability to learn novel information, the ability to recall sequences that contain repeated items (where some form of temporary marker is needed to distinguish the repetitions), and selective neuropsychological impairments of STM and LTM. Norris accepts a role for activated representations in LTM in STM tasks and this is the position we take here. Indeed, the nature of language processing forces us to consider the overlap. We assume that steps towards explaining the role of WM in language will involve understanding how transient memory for phonological sequences operates in the context of constraints that reflect the combinatorial statistics of linguistic knowledge (as we discuss in more detail later). In taking this approach, we have been encouraged by evidence implicating the phonological loop in a various aspects of language processing.

**Phonological Loop and Language Processing**

In this section, we discuss briefly some of the empirical evidence linking the phonological loop to the processes of learning new words, language production, and language comprehension. Several strands of research point to its role in vocabulary acquisition (Baddeley, Gathercole, Papagno, 1988; see also the chapter by Papagno, this volume). Most striking is the case of a neuropsychological patient with an acquired selective impairment of auditory-verbal STM who was completely unable to learn word-nonword paired-associates in which the nonwords were foreign language translations, but totally unimpaired in learning word-word pairs (Baddeley et al., 1988). This suggests a key role for the loop in maintaining novel phonological sequences where, unlike familiar words, there is no support from lexical representations in LTM. Complementing the neuropsychological evidence, experimental manipulations of articulatory suppression, phonological similarity, and word length have been shown to affect healthy adults' learning of word-nonword paired-associates much more than word-word pairs (Papagno, Valentine, Baddeley, 1991; Papagno Vallar, 1992). In children, longitudinal studies of individual differences in development suggest that, initially, vocabulary acquisition is limited by phonological loop capacity whereas at a later stage vocabulary boosts performance in STM tasks by providing lexical support (Gathercole, Willis, Emslie, Baddeley, 1992). These findings are complemented by evidence for impaired nonword repetition in children diagnosed with Specific Language Impairment (Gathercole Baddeley, 1989), Developmental Language Disorder (Graf Estes, Evans Else-Quest, 2004), and Dyslexia (Melby-Lervag Lervag, 2012). Longitudinal studies of individual differences in older children indicate a specific association between the ability to repeat nonwords and the acquisition of foreign language vocabulary over a three-year period (Service, 1992).

As might be expected, there is also strong evidence linking the phonological loop to speech production. Ellis (1980) demonstrated detailed correspondences between patterns of error in the immediate serial recall of CV and VC syllables and slips of the tongue in everyday speaking (MacKay, 1970; Nooteboom, 1973). Page, Madge, Cumming, and Norris (2007) took this further by showing that characteristic patterns of error in recalling sequences of alternating phonemically similar and dissimilar letters also occur in speech errors when such sequences are repeatedly read aloud. Interestingly, however, detailed investigation of a neuropsychological case with a marked selective impairment of auditory-verbal STM revealed that the patient's speech was completely normal (Shallice Butterworth, 1977). This would not be expected on a holistic account of the phonological loop as a speech input-output system and has been interpreted as suggesting two buffers, one for speech input the other for spoken output (Vallar Papagno, 2002), in line with an earlier proposal by Monsell (1987).

Further evidence suggests the phonological loop is also involved in language comprehension, though not crucially. Thus, for normal adults, articulatory suppression disrupts oral comprehension only when sentences are syntactically complex (Rogalsky, Matchin, Hickok, 2008). This is complemented by evidence of almost normal oral language comprehension in patients with acquired impairment of auditory-verbal STM, with problems only when sentences are unusually long and syntactically complex (Vallar Baddeley, 1987). Further research has shown the problem occurs when comprehension depends on reactivating phonological information over a distance (Friedmann Gvion, 2001). We know also that performance in verbal STM tasks is a weak predictor of individual differences in reading comprehension whereas

performance in WM span tasks that measure the capacity to combine attention-demanding mental operations with storage in STM are a much more powerful predictor (Daneman Carpenter, 1980). These observations suggest that the overall capacity of WM is more important for language comprehension than the phonological loop per se, in line with the earlier findings of Baddeley and Hitch (1974). It seems language inputs feed rapidly through to lexical and semantic streams of analysis, with the phonological loop becoming useful particularly when comprehension requires the reprocessing of verbatim information (Gvion Friedmann, 2012).

### Computational Modelling of the Phonological Loop

So far, we have described the phonological loop as a component of WM and discussed briefly some of the evidence for its involvement in three broad aspects of language processing—vocabulary acquisition, speech production and, to a more limited extent, comprehension. We have also highlighted a number of issues where the concept is in need of further development, the main ones being to incorporate effects of linguistic knowledge (LTM), the learning of new words, and differences between the "inner ear" and the "inner voice". The initial concept was deliberately parsimonious and the challenge is whether it can be elaborated to handle these and other additional effects. Most noteworthy among the latter are effects associated with maintaining the temporal order of a sequential input. The phonological loop was originally likened to a closed loop tape-recorder, but this was offered as a metaphor rather than a mechanism and cannot account for the characteristic tendency to make transposition errors in immediate serial recall where a presented item is recalled in the wrong position in the sequence. Furthermore, processing and retaining temporal order information is clearly central to language processing more generally given that this involves a structured hierarchy in which information is serially ordered both within and between levels. Thus, for example, repeating back a spoken sentence involves parallel temporal sequencing at acoustic, phonological, lexical, semantic and articulatory levels, each with their own characteristic time-scales. The need to process hierarchies of temporal order in language may be the most salient feature distinguishing verbal WM from visuo-spatial WM. As we will go on to show, we regard immediate serial recall as a tool for exploring verbal STM capacity for temporal order and as a stepping-stone towards understanding the role of transient sequential information in language processing more generally.

## Computational Models of Serial Order

As will be clear by now, *immediate serial recall* (henceforth, 'serial recall') is the dominant task used to study serial order in verbal STM. Participants are given a sequence of items (letters, digits, or words) that they must subsequently recall in the correct order. This task has generated a wealth of reproducible findings, yielding a rich set of constraints for theorising (Lewandowsky & Farrell, 2008). The wealth and richness of findings makes serial recall an ideal task for computational modelling efforts. Accordingly, several computational models have been developed that provide a quantitative account of serial recall phenomena. Some of these models are cast within the phonological loop account (Burgess & Hitch, 1999; Page & Norris, 1998), essentially providing a computational instantiation of the verbal-conceptual model that includes an explicit mechanism for serial ordering, whereas others are cast outside the phonological loop concept. The models can be broadly divided into two categories, based on the serial ordering mechanism they employ, namely chaining models and competitive queuing models. In what follows, we discuss each of these classes of models in turn with respect to their ability to explain the list of benchmark findings shown in Table 1 (for a more comprehensive list of benchmarks, see Hurlstone, 2021; Hurlstone, Hitch, & Baddeley, 2014). We invite the reader to scrutinise the benchmarks, which include key terms referred to in the text, before advancing further.

### Chaining Models

Associative chaining is the oldest approach to serial order (Ebbinghaus, 1885/1964; Kahana, 2012) and the mechanism of serial recall in several computational models (Lewandowsky & Murdock, 1989; Murdock, 1993, 1995; Solway, Murdock, & Kahana, 2012). In chaining models, serial order is encoded by forming associations between study items. Serial recall is accomplished by traversing these associations, which serve as the retrieval cues for sequence production. For example, given the study sequence *A, B, C,* retrieval of *A* will cue retrieval of *B*, which will then cue retrieval of *C*. Chaining models can be divided into two classes: simple chaining and compound chaining. In simple chaining models (Figure 1a), such as the original Theory of Distributed Associative Memory (TODAM) model (Lewandowsky & Murdock, 1989), only forward associations between adjacent items are used to represent serial order. By contrast, in compound chaining models (Figure 1b; Solway et al., 2012), which includes later instantiations of TODAM (Murdock, 1993, 1995), serial order is represented by forward and backward associations between both adjacent and non-adjacent items, the strength of which decreases gradually as a function of the distance between items, with backward associations being weaker than forward associations. An additional assumption required in chaining models is that the first item in the sequence must be associated with a start of sequence marker to kickstart the chaining process at recall.

Chaining models face several challenges. Although the chaining mechanism produces a primacy effect, it produces

Table 1
*Benchmark findings of serial recall.*

| Finding | Brief Description | References |
|---|---|---|
| Serial-position curve | When recall accuracy is plotted as a function of the serial position of items, the resulting serial-position curve exhibits a large primacy effect (superior recall of early-sequence items) and a small recency effect (enhanced recall of end of sequence items). | Drewnowski & Murdock (1980); Madigan (1971) |
| Sequence length effect | Serial recall performance decreases gradually as sequence length increases. | Crannell & Parrish (1957); Maybery et al. (2002) |
| Error types | Errors in serial recall can be transposition or item errors. Transpositions occur when items are recalled in the wrong serial positions. Item errors include intrusions (recall of prior-sequence or extra-sequence items), omissions (failure to recall an item in a position), and repetitions (items recalled on more than one occasion, despite occurring only once in the study sequence). Transpositions are typically more common than item errors. | Henson (1996) |
| Transposition gradients | The probability of transpositions decreases with increasing ordinal distance from the target position. Thus, when an item is recalled in the wrong position it will tend to be close to its correct position. This tendency for transpositions to cluster around their target positions is known as the locality constraint. | Farrell & Lewandowsky (2004); Henson et al. (1996) |
| Fill-in | If an item $i$ is recalled a position too soon, recall of item $i - 1$ is more likely at the next recall position than item $i + 1$. Thus, given the study sequence ABC, if B is recalled first, then a fill-in error, reflected by the subsequent recall of A, is more likely than an infill error, reflected by the subsequent recall of C. | Farrell et al. (2013); Surprenant et al. (2005) |
| Temporal grouping effects | Inserting an extended temporal pause after every few items in a study sequence—known as temporal grouping—improves recall accuracy, causes mini within-group primacy and recency effects, and a tendency for items to exchange groups but maintain their position within groups (a class of errors known as interpositions). | Hartley et al. (2016); Hitch et al. (1996); Ryan (1969b) |
| Phonological similarity effects | Sequences of phonologically similar sounding items (e.g., *B D G P T V*) are recalled less accurately than sequences of phonologically dissimilar sounding items (e.g., *F K L R X Y*). This phonological similarity effect is also observed when sequences are constructed by alternating phonologically dissimilar and similar items (e.g., *F B K G R T*). Such mixed sequences exhibit a saw-toothed accuracy serial position curve, with peaks corresponding to the recall of dissimilar items and troughs corresponding to the recall of similar items. | Baddeley (1968); Page et al. (2007) |
| Word length effect | Serial recall declines as the time needed to articulate items is increased. Thus, ordered recall of sequences of words with long articulation times (e.g., *coerce, harpoon, cyclone*) is worse than for sequences of words with shorter articulation times (e.g., *wicket, pectin, bishop*). | Baddeley et al. (1975)<br>but see Jalbert et al. (2011); Lewandowsky et al. (2009); Service (1998) |
| Articulatory suppression effect | Suppressing articulation by repeating an irrelevant word (e.g., "the", "the", "the"...) impairs immediate serial recall and at the same time removes the phonological similarity effect (with visual item presentation) and word length effect. | Baddeley et al. (1975); Murray (1968) |

no recency effect. The primacy effect arises because successful recall of item $i$ depends on correct recall of item $i - 1$, which in turn depends upon the correct recall of item $i - 2$. This dependency means the chaining mechanism predicts recall performance will decrease monotonically across serial positions, with performance being worst at the final position. Furthermore, although the chaining mechanism generates a primacy effect, without ancillary mechanisms the extent of primacy produced will tend to be weaker than observed empirically. To accurately model the primacy effect, Lewandowsky and Murdock (1989) had to augment TODAM with two mechanisms—a primacy gradient (see later) in the

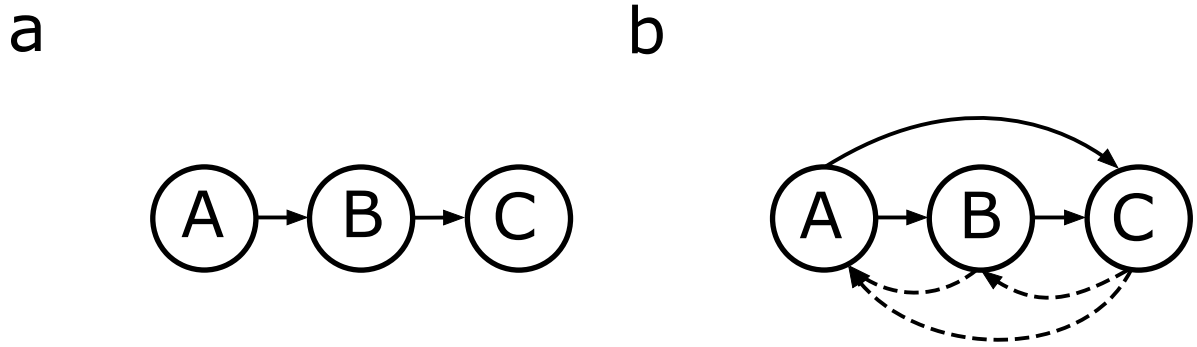


*Figure 1*. Illustration of simple (a) and compound (b) chaining models.

encoding strength of each successive association, and output interference during recall. Similarly, to explain the recency effect, two mechanisms were once again implemented—retroactive interference during the encoding of item and associative information, and response suppression (see later). Whilst independent empirical evidence can be adduced for each of these mechanisms, the combination of all four to explain two of the most basic serial recall benchmarks is hardly parsimonious. Simple chaining models also encounter difficulties explaining the locality constraint on transposition errors—since a simple chaining mechanism only activates forthcoming items, it cannot readily explain how an earlier item can take the place of a later one as an error. Compound chaining models, by contrast, can capture the pattern of transpositions by virtue of their use of bidirectional and graded associations between items (Murdock, 1995; Solway et al., 2012). Omission and intrusion errors are also problematic, since in both cases the cue for the next to-be-recalled item will have been lost, meaning serial recall must terminate before the end of the sequence is reached. This prediction is at variance with the behavioural data, since participants frequently do recover from such errors.

There are more serious objections to chaining. First, chaining accounts have difficulties explaining the pattern of findings associated with the recall of sequences in which phonologically similar and dissimilar items are intermixed (e.g., *B K P R*). Chaining accounts predict recall of the dissimilar items *K* and *R* should be impaired, because they possess similar (confusable) retrieval cues. However, this prediction is contrary to the data (Baddeley, 1968; Henson et al., 1996), which shows dissimilar items in mixed sequences are recalled as effectively as items in corresponding positions in pure dissimilar sequences. Second, chaining accounts predict more infill than fill-in errors, because an item recalled too soon will subsequently cue the item that followed it in the input sequence more strongly than any other by virtue of its direct associative link with that item. This prediction is at variance with the empirical data (Farrell et al., 2013; Surprenant et al., 2005).

Due to these limitations, and others (see Hurlstone et al., 2014), theorists have largely discounted chaining as a viable account of serial recall (although see Logan, 2021 for a recent chaining model and Osth Hurlstone, 2021 for a critique) and turned instead to an alternative class of models that we consider next.

### Competitive Queuing Models

The current dominant class of serial recall models are CQ models (Grossberg, 1978a, 1978b; Houghton, 1990). Such models were motivated by the insight that slips in performance, such as transpositions in serial recall, imply that the representation of serial order is parallel (items are co-activated simultaneously), rather than serial (one item excites the next), as is the case in chaining models (Houghton & Hartley, 1995). CQ models are characterised by a two-stage parallel-sequence-planning and response-selection mechanism, comprising an activation layer and a selection layer. In the first stage, target items are activated in parallel according to a gradient of activation by an *activating mechanism* (Glasspool, 2005) that determines the relative output priority of items. These activated representations are projected to the second stage, wherein items compete for selection through mutual inhibition and self-excitation. The item with the strongest activation level is selected for recall, after which its corresponding representation in the first stage is inhibited—an assumption known as *response suppression*. This competition is re-run until all items have been selected for output in the second stage, and their corresponding representations in the first stage have been inhibited. By adding random noise to item activation levels in the first or second stage, the CQ mechanism can simulate errors in sequence production.

The main difference between CQ models relates to the activating mechanism used to generate the activation gradient. There are two main model variants, *context-free* and *context-based*.

**Context-free CQ models.** In context-free CQ models (Figure 2a-c), the activating mechanism generates a single

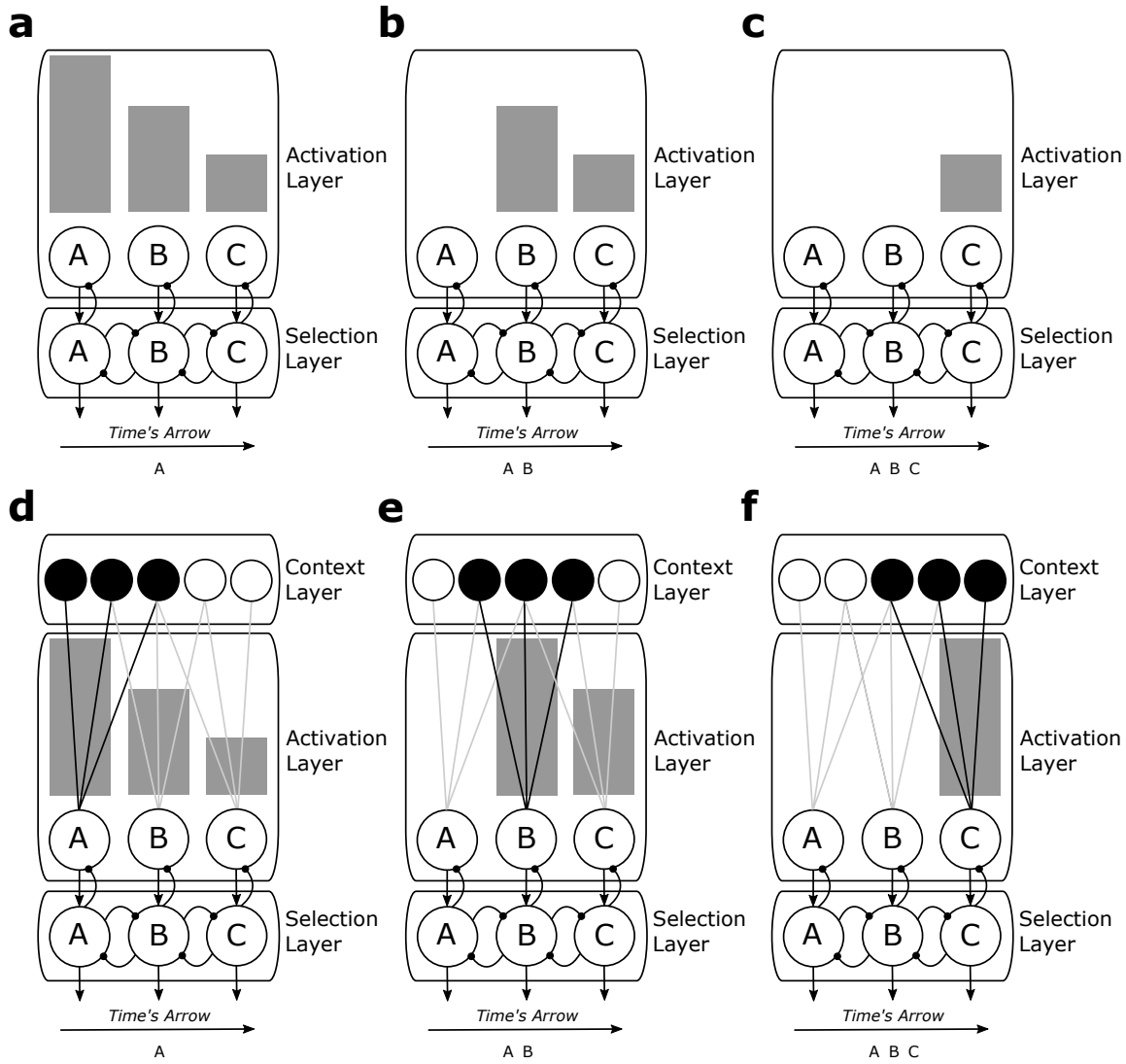


*Figure 2*. Schematic of the architecture of context-free (a-c) and context-based (d-f) CQ models and the steps involved in producing a three-item sequence. Both classes of models (all panels) comprise an activation layer and a selection layer. Lines terminating with arrows represent excitatory connections, whereas lines terminating with semicircles represent inhibitory connections. Each node in the lower selection layer has an inhibitory connection to every other node in the same layer (for simplicity only adjacent-neighbour inhibitory connections are shown) and to its corresponding node in the activation layer (the pathway through which response suppression is implemented). Columns in the activation layer represent the activation levels of the nodes representing items in the to-be-recalled sequence. In context-free CQ models, a single activation gradient—known as a primacy gradient—is established over target items in the activation layer. The activations are projected to the selection layer, wherein items compete for selection via lateral inhibition, resulting in the production of the first item and the suppression of its corresponding representation in the activation layer (a). This allows the second item to become the most active and win the response competition (b), and its suppression, in turn, allows the third item to become the most active resulting in its production (c) and subsequent suppression (not shown). In context-based CQ models, the activation gradient established over target items in the activation layer varies in response to the activity of a re-evolving context signal—originating from an additional layer, known as the context layer—to which items were temporarily bound during serial order encoding. Reinstatement of the context signal to its initial state in the context layer establishes an activation gradient over target items in the activation layer. Activations are projected to the selection layer, wherein items compete for selection, resulting in the production of the first item and the suppression of its corresponding representation in the activation layer (d). The context signal advances to the next state and a new activation gradient is established, allowing the second item to win the output competition (e), and its suppression, followed by the advancement of the context signal to its final state, generates a new activation gradient allowing the third item to become the most active, resulting in its production (f) and subsequent suppression (not shown).

activation gradient over items that is then held constant during sequence generation. This pattern of activation could originate from a short-term store, such as the phonological loop, or from LTM. The item activations are constrained by a primacy gradient, such that each item has an activation level that is weaker than its predecessor. Serial recall is accomplished via an iterative process of selecting the strongest item for recall before suppressing its activation, so the next strongest item can be selected. This is the functional mechanism for ordered recall in the models of Grossberg (1978a, 1978b), the primacy model of the phonological loop (Page & Norris, 1998), the Serial-Order-in-a-Box (SOB) model (Farrell & Lewandowsky, 2002), and the LIST PARSE model (Grossberg & Pearson, 2008). The post-output suppression of items is a crucial ingredient in these context-free CQ models because without it they would perseverate on the first response, which would always be the most active. One limitation of context-free CQ models is that the order of repeated items in a study sequence cannot be represented using type representations with a single activation level. Repeated items must therefore be handled by incorporating multiple token representations of the same item.

**Context-based CQ models.** In context-based CQ models (Figure 2d-f), each item in a sequence is associated with the current state of an internal context signal that changes gradually during encoding, and represents the positions of items in the sequence. At recall, the context signal is reset and re-evolves along its original path, with sequence items being activated according to the degree of similarity between the current state of the context signal and the state to which each item was originally associated. The context signal therefore serves as a dynamic activating mechanism that shifts the source of activation during retrieval. The dynamic nature of the activating mechanism enables these models to generate sequences containing repeated items using type rather than token representations. This is because the type representation of a repeated item will receive different sources of activation at different points during sequence generation. The post-output suppression of items is therefore less crucial in these models because the context signal carries the primary burden for sequence generation.

Context-based CQ models can be classified as event-based, time-based, or a hybrid of the two. In event-based models, the context signal changes when a new event (e.g., a new item) is experienced. In one class of event-based models, the context signal encodes absolute within-sequence position. Models falling into this class include C-SOB (Farrell, 2006; Lewandowsky & Farrell, 2008) and the original Burgess and Hitch (1992) network model of the phonological loop. For example, in the latter model, items are associated with a context signal implemented as a vector of inactive nodes containing a dynamic window of active nodes. The context vector changes gradually with the presentation of each item by sliding the moving window of activation from left to right by a constant one node per item. In another class of event-based models, the context signal encodes relative within-sequence position. For example, in the Start-End Model (SEM; Henson, 1998; see also Houghton, 1990) items are linked to the varying states of a context signal comprising two elements—a start marker that is strongest for the first position and decreases exponentially in strength across positions, and an end marker that is weakest for the first position and increases exponentially in strength across positions. Such a context signal represents approximate position relative to the start and end of the sequence.

In time-based models, the context signal changes as a function of absolute time. Models in this class include more recent instantiations of the Burgess and Hitch (1999, 2006) model in which the same moving window context signal changes with time rather than events, and the Oscillator-Based Associative Recall (OSCAR) model (Brown, Preece, & Hulme, 2000). In the OSCAR model, items are linked with the different states of a time-varying context signal driven by sets of temporal oscillators operating at different frequencies. At recall the context signal is reset to its initial state before being replayed, with sequence items being re-activated through their original associations with the timing signal. A similar, but more abstract, temporal coding scheme is utilised by the SIMPLE model (Brown, Neath, & Chater, 2007).

**CQ explanation of benchmark findings.** CQ models can account for the benchmark findings of serial recall. In context-free CQ models, the primacy effect materialises because the activation levels of items near the beginning of the sequence are more distinctive, meaning these items encounter less competition during recall than items towards the end of the sequence. By contrast, the recency effect manifests because as successive items are recalled and suppressed, the number of response competitors is gradually reduced. In context-based CQ models, primacy and recency effects are partly, if not wholly, determined by "edge effects"—there are less opportunities for items near the beginning and end of a sequence to move around, compared to items at medial positions. In some models (e.g., Brown et al., 2007; Henson, 1998), an additional contributing factor is the greater distinctiveness of the context signal at terminal positions. In both context-free and context-based CQ models, sequence length effects arise because the greater the number of items in the target sequence, the greater the probability of an error being committed. In some context-based CQ models (e.g., Henson, 1998), an additional factor contributing to the sequence length effect is that the resolution of the context signal for longer sequences is weaker than for shorter sequences. It is generally easier for CQ models to produce transpositions than item errors. This is a natural consequence of the parallel sequence dynamics assumed by these models, which when perturbed by noise will alter the relative priority of items.

Near-neighbour transpositions predominate because the representation of serial order via an activation gradient necessarily implies that the strongest competitors to the target item at each recall position will be items from adjacent, rather than remote, serial positions, thus accommodating the locality constraint. Omission errors are accommodated by incorporating an output threshold that the strongest item must exceed in order to be recalled, whilst intrusion errors are modelled by weakly activating extra-sequence items to allow them to enter into the response competition. The scarcity of erroneous repetitions is accounted for in CQ models by the post-output inhibition of items, which reduces the likelihood an item will be recalled more than once.

Context-free CQ models can accommodate the finding that fill-in errors are more frequent than infill errors. This is because if an item $i$ is recalled a position too soon and then suppressed, item $i - 1$ will be a stronger competitor at the next recall position than item $i + 1$, because the former item, by virtue of being presented earlier in the sequence, will have been encoded more strongly on the primacy gradient. Context-based CQ models can also accommodate this result (Burgess & Hitch, 1999; Henson, 1998), either by incorporating a primacy gradient as one component of the context signal (Henson, 1998), by incorporating a primacy gradient in the strength of the connections between items and the context signal (Brown et al., 2000; Lewandowsky & Farrell, 2008), or by incorporating a primacy gradient in conjunction with the context signal (Burgess & Hitch, 1999). A primacy gradient also appears to be necessary for context-based CQ models to provide an adequate account of the primacy effect, the distribution of omissions and intrusions across output positions, and the dynamics of transpositions (see Hurlstone, 2021 and Hurlstone et al., 2014 for discussion).

To accommodate temporal grouping effects, context-based CQ models assume a hierarchical or multidimensional context signal, whereby one component of the signal represents the positions of items or groups within the sequence overall, and the second component represents the positions of items within groups (Brown et al., 2000; Burgess & Hitch, 1999; Hartley, Hurlstone, & Hitch, 2016; Henson, 1998; Lewandowsky & Farrell, 2008). This provides a more distinctive two-dimensional representation of serial position, which accounts for the overall reduction in transposition errors in grouped sequences and the emergence of within-group primacy and recency effects. Interposition errors arise because items in the same positions in different groups are associated with the same state of the component of the context signal that represents within-group position, rendering them vulnerable to confusion. Grouping effects are beyond the purview of context-free CQ models, which, because of their lack of positional representations, are unable to explain the pattern of interpositions.

CQ models explain phonological confusions by assuming a third stage wherein an item chosen at the second stage undergoes a further competition in which it is vulnerable to confusion with other items based upon its degree of phonological similarity to those items (Burgess & Hitch, 1999; Henson, 1998; Page & Norris, 1998). The effect of this is to increase the likelihood that a similar item recalled from the second stage will be confused with another similar item in the third stage, thus accommodating the poorer ordered recall of phonologically similar compared to dissimilar items in pure and mixed sequences.

Two CQ models of the phonological loop, namely the primacy model (Page & Norris, 1998) and the Burgess and Hitch (1999) model have been applied to the word length effect. Both models assume that forgetting in short-term memory is attributable to trace decay. Thus, in the primacy model the primacy gradient of activations held over items decays gradually with the passage of time but can be refreshed by a process of cumulative rehearsal. In the Burgess and Hitch (1999) model, such trace decay occurs in the connections between items and the context signal used to represent serial order, as well as connections between items and input and output phonemes that implement the phonological loop. Rehearsal involves an iterative process whereby items activate output phonemes corresponding to their pronunciation pattern. The output phonemes then send activation to input phonemes, which then send activation back to the items. On each cycle of rehearsal, the strength of connections between items and output phonemes and input phonemes and items are partially strengthened. In keeping with Baddeley's verbal conceptual account, both models account for the word length effect because long words take longer to articulate then shorter words, meaning there is less opportunity to engage in rehearsal to offset the (item or weight) decay process.

The model of Burgess and Hitch (1999) can also simulate the articulatory suppression effect. The effect is approximated by adding noise to the output phonemes corresponding to the spoken pronunciation of items. This noise propagates to the input phonemes and through to the items, producing interference and blocking the use of the phonological loop for rehearsal. This causes a reduction in serial recall performance, although the degree of disruption is much more pronounced than observed empirically. Through this mechanism, Burgess and Hitch (1999) are able to simulate the abolition of the word length effect and phonological similarity effect (with visual item presentation) under articulatory suppression.

This section has focused on computational modelling of serial order at the lexical level but has ignored the constraints from language knowledge, new word representation and learning we discussed at the outset. We turn next to computational approaches to serial order that take these linguistic constraints into account.

### Linguistically Constrained Computational Models

When considering the adequacy of computational models to account for phenomena in language and WM, it is necessary to go beyond the question of how arbitrary sequences of familiar items are transiently stored and recalled and begin to consider constraints that arise from the structure of language. In natural language, the sequence of phonemes in a syllable, morphemes in a word, or words in a sentence are each subject to constraints that make some sequences impossible, ungrammatical, or meaningless, while others vary hugely in their probability as determined by both perceptual and motor limitations and by long-term linguistic knowledge as reflected in phonetics, phonology, semantics, syntax, and so on. While computational models of serial order in STM provide a useful starting point, there is plainly a very long way to go in understanding how these influences contribute to memory for real-world language and the role of WM in comprehension, production, and communication. Working memory researchers, psycholinguists, and others with an interest in the mechanisms of verbal memory each approach these questions with different goals and priorities. Can the best insights of models of language, serial order, and WM be reconciled to arrive at a consensus theory, and if so how can such theoretical integration be achieved? In attempting to integrate different models it is important to consider the theoretical territory each model occupies. Where is a given model situated and how does it relate to other models, the empirical effects it explains, and those it does not address or with which it conflicts?

### Central Role of Serial Order in Theoretical Integration

We think about these questions in terms of a simplified and idealized "theoretical space" described by "dimensions" of scope and abstraction, which we can view from different perspectives within which certain empirical distinctions become more or less important.

"Scope" (illustrated on the *x*-axis in Figure 3) refers to the range of empirical phenomena addressed by the model, the experimental results that it can simulate including existing observations and predictions. Models with broader scope typically explain different phenomena through more abstract mechanisms (for example, in language models mechanisms that are independent of modality or task), whereas models that have narrower scope can achieve greater granularity in their explanations of phenomena by incorporating specialized mechanisms (for example, specific to orthography, audition, or motor processing). To characterize this difference, we can think of an additional dimension "abstraction" (illustrated on the *y*-axis in Figure 3), which is perhaps loosely linked to Marr's concept of levels of analysis (Marr, 1982).

Each implemented model tends not to cover the entire space, incorporating both fine grain (e.g., modality-specific) mechanisms and general shared principles. This is because the explanatory value of general principles would often be undermined by more detailed specification, while the explanatory value of implementational detail is often necessarily limited to a specific domain. In trying to integrate models, we are looking to identify a coherent chain of models whose overlapping scope and abstraction can connect abstract general models to very specific and detailed predictions about empirical phenomena.

In reality however, scope is a multidimensional space. In the current context, we are considering computational models of WM for language, which leads us to focus on a specific range of relevant empirical phenomena as outlined above, but even within this area different scientists may take different views about the most important issues. For example, psycholinguists may favour models and explanatory chains that extend to and connect with linguistic phenomena beyond WM, while cognitive psychologists might prefer models that connect with more general WM phenomena (such as amodal attention) that are less immediately relevant to language. This means there are multiple valid "perspectives" on the same theoretical space.

In the current context (as illustrated in Figure 3), mechanisms of serial order are central to both language and WM. Thus, they play a central role in connecting detailed implementations of verbal WM in specific tasks (such as learning novel phonological forms, digit span, or sentence production) with more abstract and general theories (such as Baddeley and Hitch's (1994) multicomponent model of WM or Gupta and MacWhinney's (1997) account of vocabulary acquisition). For this reason, some of the most promising avenues for theoretical integration lie in reconciling models of serial representation and processing, and in identifying common principles underpinning the serial representation and processing of linguistic content. Here we can identify a degree of consensus in the computational principles identified above. In particular, CQ models implicate context signals in the rapid acquisition of new unfamiliar sequences. Moreover, because all new words are unfamiliar the first time they are encountered, these principles are likely to play an important role in constraining the development of long-term lexical-phonological representations.

### Context Signals Incorporating Linguistic Constraints

In the context of language, these time-varying signals may need to be sensitive to linguistic structure if they are to account for linguistic constraints on errors in immediate recall (i.e., in STM) and spontaneous speech (i.e., in the retrieval of long-term phonological representations). Errors in both situations may be constrained by similar linguistic factors (see e.g., Boomer & Laver, 1968; Dell & Reich, 1981; Ellis, 1980; Shattuck-Huffnagel, 1979; Treiman & Danis, 1988; Vousden, Brown & Harley, 2000). For example,

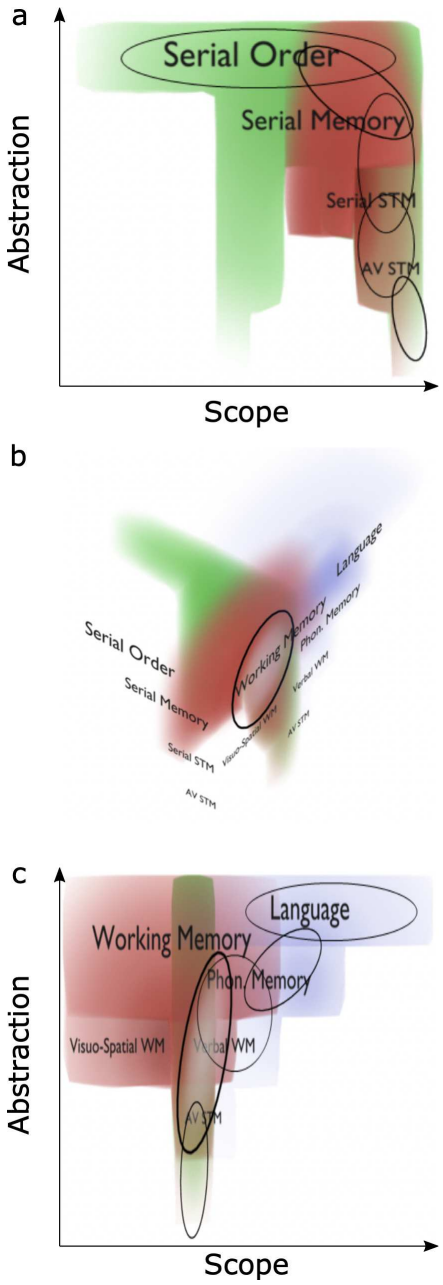


*Figure 3*. Diagram illustrating how serial order is central in connecting detailed implementations of working memory in specific tasks with more abstract and general theories. (a) we can regard models (black ellipses) as occupying a position in a space defined by scope (empirical phenomena they address) and abstraction (the degree to which they characterize general principles). An abstract model (such as the concept of CQ) can provide general principles that span and connect theories of WM (red) and serial order (green). But more detailed and less abstract models are needed to address specific forms of serial memory (such as auditory-verbal STM). Models that overlap in terms of scope and abstraction can potentially form an explanatory chain, connecting abstract general principles to concrete specific empirical phenomena. (b) however, the theoretical space has many dimensions of scope. For example, language (blue) encompasses phenomena that extend well beyond WM, and phonological memory includes long-term lexical-phonological knowledge. (c) viewed from a psycholinguistic perspective, a different chain of models may be needed to connect the general principles of language to those of WM, but the mechanisms of serial order are likely to play a central role.

phonological errors tend to involve paired or partial transpositions of phonemic units between corresponding parts of the same syllable. This contrasts with order errors in unstructured sequences (e.g., as used in digit span tasks) which tend to involve transpositions between adjacent serial positions, but bears a resemblance to the interposition errors observed when such sequences are temporally grouped (see Table 1). In the CQ modelling framework, syllabic constraints on phonological errors might be explained by a context signal that is sensitive to syllable structure, so that phonemes associated with corresponding parts of nearby syllables are associated with similar states of the context signal. For example, in a spoonerism such as "barn door" → "darn bore" the syllable initial /d/ and /b/ would be associated with similar states of the context signal, such that they compete with one another to be output at the initial position in each syllable. If the wrong phoneme is selected for output in the first syllable, it is suppressed leaving the unselected initial consonant free to be selected at the initial position in the second syllable, but other phonemes (e.g., vowels, syllable-final consonants) do not compete because they are associated with distinct states of the context signal.

Although the spoonerism example reflects constraints seen in spontaneous speech errors, as noted above a very similar pattern is seen in nonword repetition (e.g., Ellis, 1980; Treiman & Danis, 1988). In the context of this STM task, however, it becomes clear that any linguistic structure present in the context signal must derive from the unfamiliar sequence itself, because different syllable structures are subject to different patterns of error—each tending to allow transposition only to corresponding parts of different syllables. This implies, in the CQ framework, that some process must extract a representation of syllable structure online (as new words are encountered) to shape the context signal. An additional challenge—especially if we consider the learning of new phonological forms in the context of language development—is that we cannot assume that the process taps into a pre-existing representation of syllable structure, rather it must rely on cues present in the stimulus itself as it is first encountered. In other words, new word learning seems to demand a bottom-up representation of syllable structure, rather than one based on top-down application of pre-existing linguistic knowledge.

To resolve these problems, Hartley and Houghton (1996) proposed a CQ model for phonological sequencing in auditory verbal STM that incorporated a cyclical context signal sensitive to the sonority of successive phonemes. Sonority is a linguistic property of the phoneme corresponding to its relative loudness (for instance vowels are more sonorous than voiceless fricatives). Because of the sonority principle (Selkirk, 1984), phonemes in well-formed syllables typically conform to a pattern in which increasingly sonorous consonants build toward a peak (vowel) after which sonority declines over successive segments. This creates a wave-like pattern of peaks and troughs, and the phase of the cycle can be used to associate phonemes with the appropriate part of each syllable, accounting for the constraints seen in nonword repetition errors.

Although the original Hartley and Houghton model was based on the concept of sonority, it can be implemented using acoustic cues in the envelope of the speech signal (more sonorous phonemes coincide with peaks in the envelope for voiced frequencies, leading to a quasi-periodic amplitude modulation at the speech rate). Hartley (2002) constructed a syllable tracking context signal by combining the outputs of a number of oscillators processing and tuned to different frequencies of amplitude modulation spanning the range of typical speech rates. The syllabic phase model shows how auditory processing can track the speech rate, yielding—through a bottom-up process—a context signal that is sensitive to within-syllable position. This is illustrated in Figure 4, which shows the responses of oscillators in the syllabic phase model during processing of the spoken sentence "Iguanas and alligators are tropical." Panel (a) shows the raw speech signal, whereas panels (b) and (c), respectively, show the combined phase and amplitude responses of the syllabic phase oscillators over time. Syllable boundaries correspond with troughs in the envelope in panel (b) and it can be seen by comparison with panel (a) that these boundaries occur in phonologically plausible locations in the majority of cases. The syllabic phase model is consistent with parallel discoveries showing that activity in auditory cortex tracks the speech signal when participants hear natural connected speech (e.g., Ahissar et al., 2001, see Ding & Simon, 2014 for review).

## Temporal Grouping in Working Memory: A Link to Prosody?

It is natural to ask whether the same kind of mechanism might apply to larger scale serial structure in language, for example governing the order of words in a sentence. However, natural sentences have a very complex hierarchical structure which potentially combines influences of syntax, semantics, and so on. Again, the serial recall literature from laboratory studies provides useful constraints. Even in the absence of top-down information when recalling arbitrarily ordered sequences of words, as noted above, performance is influenced by the timing of the items. Regular grouping is advantageous for immediate recall and leads to characteristic changes in the distribution of order errors. The serial position curve shows multiple-bowing reflecting primacy and recency effects within groups, adjacent transposition errors are reduced, while transpositions between corresponding items in different groups increase. The latter effect is strikingly similar to the pattern seen in syllable-level transpositions between phonemes addressed in the Hartley and Houghton (1996) and Hartley (2002) models. Could similar mechanisms, reflect-

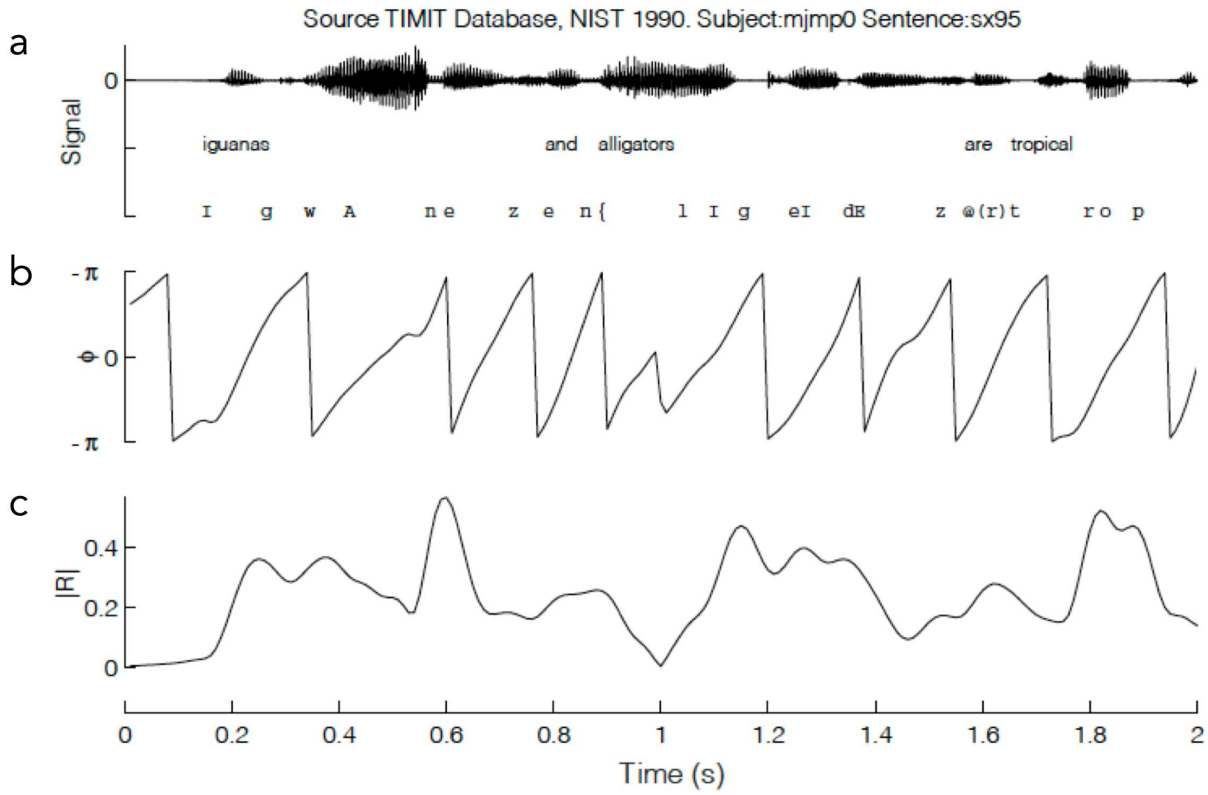


*Figure 4*. Example of responses of syllabic phase model oscillators during processing of the sentence "Iguanas and alligators are tropical". (a) shows the raw speech signal being processed. Phonetic and orthographic annotations show the locations of the phonemes and words. (b) and (c) shows how the phase and amplitude responses, respectively, of oscillators varies over time.

ing prosodic and rhythmic structure, be at work in the sequencing of words?

The regular grouping of sequences in serial recall experiments is indeed rather artificial, and it might be argued that grouping effects can be explained in terms of top-down expectations arising from the repeated and predictable pattern used in STM experiments. However, Ryan (1969a) had shown that when the sequence structure was varied (so that group sizes were irregular) similar—but necessarily more complex patterns emerged with local recency and primacy. Our own work (Hartley et al., 2016) showed that these patterns were reliable and reproducible even when the grouping pattern was varied on a trial-by-trial basis, so participants could have no foreknowledge of the sequence structure.

We were able to explain these patterns using a CQ model with a context signal comprising a bottom-up multi-scale population (BUMP, also described in Hartley et al., 2016) of oscillators similar to those used by Hartley (2002). Again, the oscillators are sensitive to local amplitude modulations in the envelope of incoming speech, with each oscillator possessing an intrinsic tuning—a tendency for its activity to oscillate at a specific rate. The frequency tunings of the oscillators are chosen to span the range of presentation rates encountered in spontaneous speech. Some oscillators are sensitive to slow modulations on a temporal scale corresponding to the length of the sequence (say 5 s), others are sensitive to more rapid fluctuations corresponding to groups of items (say 1-2 s), and yet others are sensitive to faster modulations corresponding to presentation of the individual items (0.75 s). In the BUMP model, when a sequence is presented, the rhythm and timing of items determines which oscillators become entrained to the bottom-up input. The context signal is thus similar to that suggested in OSCAR (Brown et al., 2000), but unlike that model, BUMP incorporates a bottom-up mechanism explaining how the context signal arises and evolves in response to irregular and unpredictable speech.

To illustrate, in the case of an evenly-timed, ungrouped sequence, oscillators with tunings close to the item presentation rate will respond strongly and in phase with the items (Figure 5a). These oscillators will go through one cycle per sequence item. Oscillators with tunings close to the sequence presentation rate respond to the larger scale amplitude fluc-

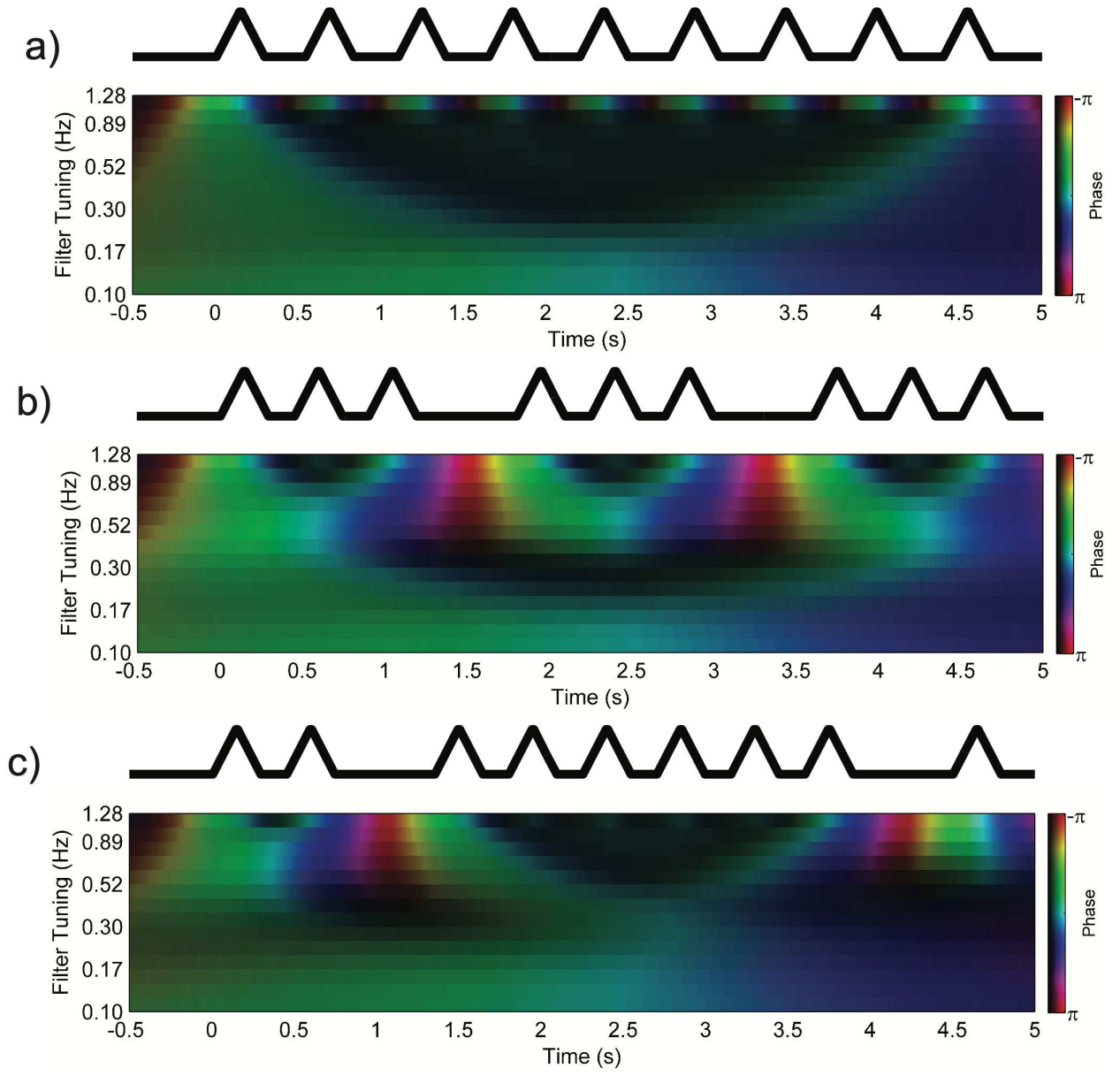


*Figure 5.* Phase and amplitude responses of a population of oscillators with different tunings (spaced between 0.1 Hz and 1.28 Hz) in the BUMP model of Hartley et al. (2016). For each axis, the upper trace shows the timing of triangular amplitude pulses representing each item in a nine-item sequence. The coloured phase amplitude diagram represents the evolution of amplitude and phase of oscillators with different tunings (*y*-axis) over time (*x*-axis). Phase is indicated by hue, amplitude by brightness. (a) Ungrouped sequence, (b) sequence grouped into three groups of three (3-3-3), (c) 2-6-1 grouped sequence. Note that the same population of oscillators responds to all three sample sequences, but the phase and amplitude of the entrained responses is systematically affected by the sequence structure and in particular by local amplitude modulations on different scales (e.g., corresponding to sequence, group, item). Each item will be associated with the state of the oscillator population at the time it is presented. At retrieval items associated with similar states may be confused with one another, resulting in transposition errors.

tuation associated with presentation of the entire sequence and are insensitive to the relatively rapid changes associated with individual items. These slower oscillators' output will go through approximately half a cycle during presentation of the sequence. Oscillators with intermediate tunings respond only weakly and their responses are largely restricted to the beginning and end of the sequence. However, for a regularly grouped sequence, oscillators with tunings close to the group presentation rate are also recruited, and go through one cycle per group (Figure 5b). Oscillators with tunings close to the group presentation rate are also recruited for irregularly grouped sequences, but the inconsistency of group durations means oscillators in this range are less strongly activated than would be the case for regularly grouped sequences (Figure 5c). An important property of the model is that some temporal grouping patterns are more favourable than others. Regular grouping patterns powerfully activate oscillators tuned to the grouping rate, which enhances overall recall, albeit at the expense of interposition errors (as seen in the similarity of the outputs of filters tuned to the group presentation rate in Figure 5b). Irregular grouping patterns similarly favour intergroup transpositions although the correspondence between different positions is less clear-cut (see Figure 5c).

In addition to explaining the pattern of errors based on the sequence structure, the BUMP model overcomes several theoretical barriers to the wider application of CQ in language processing. Specifically, it provides a mechanism that avoids two formerly unexplained problems that affected the capacity of earlier models to deal with more realistic verbal sequences seen in natural language: (1) how to anticipate the start and end of a sequence, and (2) how to choose the appropriate rate of change of the context signal. BUMP is a hybrid model—the context signal changes smoothly over time, but these changes are driven by events.

### Wider Considerations

In our own work and in the models reviewed above, we have focused mainly on putatively hard-wired or bottom-up mechanisms that can rapidly learn and retrieve unfamiliar sequences in the absence of relevant long-term knowledge. We argue that these mechanisms must play a foundational role in language because STM for unfamiliar verbal sequences is a necessary precursor to long-term linguistic knowledge. However, it is clearly not the end of the story, and even in laboratory tasks it is clear that long-term knowledge can influence memory for items and their serial order, as seen in semantic grouping effects in word-sequence recall (Kowialiewski et al., in press) and chunking in sentence recall (Baddeley, Hitch Allen, 2009). Thus, the bottom-up approach will only take us so far, even if we confine ourselves to the laboratory.

CQ remains the most promising theoretical mechanism connecting and potentially unifying theories of WM and serial order (Figure 3a) with related mechanisms in language (Figure 3b & 3c). Some progress has been made toward understanding the mechanisms of sequencing at phonological and lexical levels, albeit largely focusing on laboratory tasks such as serial recall and nonword repetition. Further progress is likely to require consideration of semantic, syntactic, pragmatic, and prosodic factors that potentially modulate these mechanisms, and which will likely require a much richer and more realistic account of linguistic representation.

Recent developments in machine learning indicate that such models will demand a greater role for long-term statistical structure in the serial processes underpinning language production and comprehension—emergent top-down knowledge that can only be acquired with substantial experience (McClelland et al., 2011; Schwering & MacDonald, 2020). We earlier argued that "chaining" models could be ruled out in explaining serial order in working memory, because of the many problems they face in accounting for ordering errors in serial recall. It has also been argued (Houghton & Hartley, 1995) that chaining leads to unavoidable interference between sequences stored in LTM (for example, between the phonological sequences that comprise familiar words with overlapping subsequences). However, more recent work has demonstrated that, with extensive training, recurrent neural networks (Jordan, 1986; Elman, 1990; Hochreiter & Schmidhuber, 1997)—in some ways resembling compound chaining models—can overcome these limitations, developing rich representations that capture serial structure in their training material and can support STM for serial order (Botvinick & Plaut, 2006; intriguingly the trained STM model shows CQ-like dynamics). Building on these approaches, "large language models", more recent recurrent architectures with billions of interconnections and trained on huge text corpora, are capable of generating remarkably realistic and meaningful linguistic output incorporating a wide variety of high-level linguistic constraints (Radford et al., 2019, GPT-2; Brown et al. 2020, GPT-3). These developments are beyond the scope of the current article but they suggest that a full account of the role of WM in language may require a more detailed and realistic implementation of long-term linguistic knowledge, based on the statistical properties of language, and a big challenge for the future is to reconcile these mechanisms with the general mechanisms of verbal WM we have focused on here.

## Conclusion

We began by reviewing evidence for a phonological loop in verbal WM involved in immediate serial recall and aspects of language processing that include lexical acquisition, speech production, and comprehension. Overall, our review of computational models of immediate serial recall indicates a central role for serial order in both WM and language. Verbal WM demands a mechanism that can encode novel se-

quences rapidly, and computational models of this process highlight the importance of CQ, which involves parallel sequence planning and competitive response selection. To account for linguistic constraints on nonword repetition errors and effects of rhythm in serial recall we argue that this mechanism must also involve a time varying context signal that is sensitive to linguistic structure. We focused on low-level auditory-temporal structure that is most relevant to STM, but we acknowledge that a fuller account of linguistic structure will involve representations that incorporate the high-level statistical constraints that arise from syntax and semantics but which can only be learned over long-term experience. Language models and WM models have thus tended to occupy different levels of abstraction and empirical scope reflecting the distinct perspectives of psychologists, linguists, cognitive scientists, and, increasingly, AI developers. By targeting the remaining gaps in this space, such as the disjunction between mechanisms of sequencing in large language models and in models of STM, we believe modelling serial order and WM can play a vital role in understanding language processing.

## References


Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences, 98(23),* 13367-13372.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Sppence (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 2, pp. 89-195). New York: Academic Press.

Baddeley, A. D. (1966a). Influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology, 18,* 302-309.

Baddeley, A. D. (1966b). 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? *The Quarterly Journal of Experimental Psychology, 20,* 249-264.

Baddeley, A. D. (1986). *Working Memory*. Oxford, UK: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4,* 417-423.

Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia, 49,* 1393-1400.

Baddeley, A., Eldridge, M., & Lewis, V. (1981). The role of subvocalization in reading. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 33,* 439-454.

Baddeley, A.D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Baddeley, A.D. & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. VIII, pp. 47-90). New York: Academic Press.

Baddeley, A. D., Hitch, G. J., & Allen, R.J. (2009). Working memory and binding in senntence recall. *Journal of Memory and Language, 61,* 438-456.

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

Baddeley, A.D., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language, 27,* 586-595.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and structure of short-term-memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575-589.

Baddeley, A. D. & Warrington, E. K. (1970). Amnesia and distinction between long-and short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 176-189.

Besner, D., Davies, J., & Daniels, S. (1981). Reading for meaning - the effects of concurrent articulation. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 33,* 415-437.

Boomer, D. S. & Laver, J. D. M. (1968). Slips of the tongue. *British Journal of Disorders of Communication, 3,* 2-12.

Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology, 26,* 467-482.

Broadbent, D. E. (1958). *Perception and Communication*. New York: Pergamon Press.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114,* 539-576.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107,* 127-181.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Burgess, N. & Hitch, G. J. (1992). Towards a network model of the articulatory loop. *Journal of Memory and Lan-*

*guage, 31,* 429-460.

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

Burgess, N. & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language 55,* 627-652.

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

Cowan, N. (1995). *Attention and Memory: An Integrated Framework*. Oxford Psychology Series, No. 26. New York, NY: Oxford University Press.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Crannell, C. W. & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters, and words. *The Journal of Psychology, 44,* 319-327.

Crowder, R. G. (1982). The demise of short-term-memory. *Acta Psychologica, 50,* 291-323.

Daneman, M. & Carpenter, P. A. (1980). Individual-differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450-466.

Dell, G. S. & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20(6),* 611-629.

Ding, N. & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience, 8,* 311.

Drewnowski, A. & Murdock, B. B., Jr. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6,* 319-332.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology*. New York: Dover.

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.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Farrell, S. (2006). Mixed-list phonological similarity effects in delayed serial recall. *Journal of Memory and Language, 55,* 587-600.

Farrell, S., Hurlstone, M. J., & Lewandowsky, S. (2013). Sequential dependencies in recall of sequences: Filling in the blanks. *Memory & Cognition, 41,* 938-952.

Farrell, S. & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9,* 59-79.

Farrell, S. & Lewandowsky, S. (2004). Modelling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language, 51,* 115-135.

Friedmann, N. & Gvion, A. (2001). Sentence comprehension and working memory limitation in aphasia: A dissociation between semantic-syntactic and phonological reactivation. *Brain and Language, 86(1),* 23-39.

Gathercole, S.E. & Baddeley, A.D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28,* 200-213.

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology-Learning Memory and Cognition, 25,* 84-95.

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years - A longitudinal-study. *Developmental Psychology, 28,* 887-898.

Glanzer, M. & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5,* 351-360.

Glasspool, D. W. (2005). Modelling serial order in behaviour: Evidence from performance slips. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 241-270). Hove: Psychology Press.

Graf Estes, K., Evans, J.L., & Else-Quest, N.M. (2004). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech and Hearing Research, 50,* 177-195.

Gregg, V. H., Freedman, C. M., & Smith, D. K. (1989). Word-frequency, articulatory suppression and memory span. *British Journal of Psychology, 80,* 363-374.

Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & Snell (Eds.), *Progress in theoretical biology* (Vol. 5, pp. 233-374). New York: Academic Press.

Grossberg, S. (1978b). Behavioral contrast in short-term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology, 17,* 199-219.

Grossberg, S. & Pearson, L. R. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: Towards a unified theory of how the cerebral cortex works. *Psychological Review, 115,* 677-732.

Gupta, P. & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59(2),* 267-333.

Gvion, A. & Friedmann, N. (2012). Does phonological working memory impairment affect sentence compre-

hension? A study of conduction aphasia. *Aphasiology, 26(3-4),* 494-535.

Hartley, T. (2002). Syllabic phase: a bottom-up representation of the temporal structure of speech. In J. Bullinaria & W. Lowe (Eds.), *Connectionist Models of Cognition and Perception* (pp. 277-288). Singapore, World Scientific Publishing Co.

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

Hartley, T., Hurlstone, M. J., & Hitch, G. J. (2016). Effects of rhythm on memory for spoken sequences: A model and tests of its stimulus-driven mechanism. *Cognitive Psychology, 87,* 135-178.

Henson, R. N. A. (1996). Short-term memory for serial order (Doctoral dissertation). University of Cambridge, Cambridge, United Kingdom.

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, 49A,* 80-115.

Hitch, G. J. (1978). Role of short-term working memory in mental arithmetic. *Cognitive Psychology, 10,* 302-323.

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology, 49A,* 116-139.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8),* 1735-1780.

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* (pp. 287-319). London: Academic Press.

Houghton, G. & Hartley, T. (1995). Parallel models of serial behavior: Lashley revisited. *Psyche, 2(25),* 1-25.

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words - evidence for a long-term-memory contribution to short-term-memory span. *Journal of Memory and Language, 30,* 685-701.

Hurlstone, M. J. (2021). Serial recall. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of Human Memory*. Oxford University Press.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin, 140,* 339-373.

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology-Learning Memory and Cognition, 37,* 338-353.

Jordan, M. I. (1986). Serial order: a parallel distributed processing approach. Finding structure in time. *Cognitive Science, 14,* 179-211.

Kahana, M. J. (2012). *Foundations of Human Memory*. New York: Oxford University Press.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189-217.

Kowialiewski, B., Gorin, S., & Majerus, S. (in press). Semantic knowledge constrains the processing of serial order information in Working Memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Lewandowsky, S. & Farrell, S. (2008). Short-term memory: New data and a model. *The Psychology of Learning and Motivation, 49,* 1-48.

Lewandowsky, S. & Murdock, B. B. (1989). Memory for serial order. *Psychological Review, 96,* 25-57.

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences, 13,* 120-126.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10,* 447-454.

Logan, G. D. (2021). Serial order in perception, memory, and action. *Psychological Review, 128(1),* 1-44.

Logie, R. H. (1995). *Visuo-spatial Working Memory*. Hove, UK: Erlbaum.

Logie, R. H., Camos, V., & Cowan, N. (2021). *Working Memory: State of the Science*. Oxford University Press.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14(8),* 348-356.

MacDonald, M.E. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science, 25(1),* 47-53.

Mackay, D. G. (1970). Spoonerisms: Structure of errors in serial order of speech. *Neuropsychologia, 8,* 323-350.

Madigan, S. A. (1971). Modality and recall order interactions in short-term memory for serial order. *Journal of Experimental Psychology, 87,* 294-296.

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco, Freeman & Co.

Martin, N. & Saffran, E. M. (1997). Language and auditory-verbal short-term memory impairments: Evidence for common underlying processes. *Cognitive Neuropsychology, 14,* 641-682.

Martin, R. C. & Slevc, L. R. (2014). Language Production and Working Memory. In M. Goldrick, V. Ferreira & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 437-450). New York, NY: Oxford University Press.

Maybery, M. T., Parmentier, F. B. R., & Jones, D. M. (2002). Grouping of list items reflected in the timing of recall: Implications for models of serial verbal memory. *Journal of Memory and Language, 47,* 360-385.

Melby-Lervag, M. & Lervag, A. (2012). Oral language skills moderate nonword repetition skills in children with dyslexia: A meta-analysis of the role of nonword repetition skills in dyslexia. *Scientific Studies of Reading, 16,* 1-34.

Miller, G. A. (1956). The magical number 7, plus or minus 2: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Monsell, S. (1987). On the relation between lexical input and output pathways for speech. In A. Allport, D. G. MacKay, W. Prinz & E. Scheerer (Eds.), *Cognitive Science Series. Language Perception and Production: Relationships Between Listening, Speaking, Reading and Writing* (pp. 273-311). London: Academic Press.

Morey, C. C. (2018). The Case Against Specialized Visual-Spatial Short-Term Memory. *Psychological Bulletin, 144,* 849-883.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100,* 183-203.

Murdock, B. B. (1995). Developing TODAM: Three models for serial order information. *Memory & Cognition, 23,* 631-645.

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology, 78,* 679-684.

Nelson, T.O. & Rothbart, R. (1972). Acoustic savings for items forgottten from long-term memory. *Journal of Experimental Psychology, 93(2),* 357-360.

Nooteboom, S. G. (1973). The tongue slips into patterns. In V. A. Fromkin (Ed.), *Speech Errors as Linguistic Evidence* (pp. 144-156). The Hague, NL: Mouton.

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin, 143,* 992-1009.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning Memory and Cognition, 28,* 411-421.

Osth, A. F., & Hurlstone, M. J. (in press, 2021). Do item-based context representations underlie serial order in cognition? Commentary on Logan (2021). *Psychological Review*.

Page, M. P. A., Madge, A., Cumming, N., & Norris, D. G. (2007). Speech errors and the phonological similarity effect in short-term memory: Evidence suggesting a common locus. *Journal of Memory and Language, 56,* 49-64.

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

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language, 30,* 331-347.

Papagno, C., & Vallar, G. (1992). Phonological short-term-memory and the learning of novel words: The effect of phonological similarity and item length. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology, 44,* 47-67.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1(8),* 9.

Rogalsky, C., Matchin, W., & Hickok, G. (2008). Broca's area, sentence comprehension, and working memory: an fMRI study. *Frontiers in Human Neuroscience, 2:* 14. doi: 10.3389/neuro.09.014.2008

Ryan, J. (1969a). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology, 21,* 137-147.

Ryan, J. (1969b). Temporal grouping rehearsal and short-term memory. *Quarterly Journal of Experimental Psychology, 21,* 148-155.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition, 21,* 168-175.

Schwering, S. C. & MacDonald, M. C. (2020). Verbal working memory as emergent from language comprehension and production. *Frontiers in Human Neuroscience, 14:* 68. doi: 10.3389/fnhum.2020.00068

Selkirk, E. (1984). On the major class features and syllable theory. In M. Aronoff & R. T. Orehrle (Eds.), *Language Sound Structure: Studies in Phonology Presented to Morris Halle by his Teachers and Students*. MIT press.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 45,* 21-50.

Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 51,* 283-304.

Shallice, T. & Butterworth, B. (1977). Short-term-memory impairment and spontaneous speech. *Neuropsychologia, 15,* 729-735.

Shallice, T. & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology, 22,* 261-273.

Shattuck-Huffnagel, S. (1979) Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E Cooper & E. C. T Walker, (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Erlbaum, Hillsdale, N. J.

Shulman, H.G. (1972). Encoding and retention of semantic and phonemic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 499-508.

Solway, A., Murdock, B. B., & Kahana, M. J. (2012). Positional and temporal clustering in serial order memory. *Memory & Cognition, 40,* 177-190.

Surprenant, A. M., Kelley, M. R., Farley, L. A., & Neath, I. (2005). Fill-in and infill errors in order memory. *Memory, 13,* 267-273.

Treiman, R. & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(1),* 145.

Vallar, G. & Baddeley, A. (1987). Phonological short-term store and sentence processing. *Cognitive Neuropsychology, 4,* 417-438.

Vallar, G. & Papagno, C. (2002). Neuropsychological Impairments of Verbal Short-term Memory. In A. D. Baddeley & M. D. Kopelman & B. A. Wilson (Eds.), *Handbook of Memory Disorders* (2nd. ed., pp. 249-270). Chichester, UK: Wiley.

Vousden, J. I., Brown, G. D., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology, 41(2),* 101-175.

Yang, T. X., Allen, R. J., & Gathercole, S. E. (2016). Examining the role of working memory resources in following spoken instructions. *Journal of Cognitive Psychology, 28,* 186-198.

Ylinen, S., Nora, A., Leminen, A., Hakala, T., Huotilainen, M., Shtyrov, Y., MÃd'kelÃd', J.P., & Service, E. (2015). Two distinct auditory-motor circuits for monitoring speech production as revealed by content-specific suppression of auditory cortex. *Cerebral Cortex, 25(1),* 1576-1586.