Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-024-02608-y

THEORETICAL/REVIEW

# Simple Recurrent Networks are Interactive

**James S. Magnuson**[1,2,3] · **Sahil Luthra**[4]

Accepted: 19 October 2024


**Abstract**
There is disagreement among cognitive scientists as to whether a key computational framework – the Simple Recurrent Network (SRN; Elman, *Machine Learning, 7*(2), 195–225, 1991; Elman, *Cognitive Science, 14*(2), 179–211, 1990) – is a feedforward system. SRNs have been essential tools in advancing theories of learning, development, and processing in cognitive science for more than three decades. If SRNs were feedforward systems, there would be pervasive theoretical implications: Anything an SRN can do would therefore be explainable without interaction (feedback). However, despite claims that SRNs (and by extension recurrent neural networks more generally) are feedforward (Norris, 1993), this is not the case. Feedforward networks by definition are acyclic graphs – they contain no loops. SRNs contain loops – from hidden units back to hidden units with a time delay – and are therefore cyclic graphs. As we demonstrate, they are *interactive* in the sense normally implied for networks with feedback connections between layers: In an SRN, bottom-up inputs are inextricably mixed with previous model-internal computations. Inputs are transmitted to hidden units by multiplying them by input-to-hidden weights. However, hidden units simultaneously receive their own previous activations as input via hidden-to-hidden connections with a one-step time delay (typically via context units). These are added to the input-to-hidden values, and the sums are transformed by an activation function. Thus, bottom-up inputs are mixed with the products of potentially many preceding transformations of inputs and model-internal states. We discuss theoretical implications through a key example from psycholinguistics where the status of SRNs as feedforward or interactive has crucial ramifications.

**Keywords** Interaction · Neural networks

## Introduction

Simple Recurrent Networks (SRNS; Elman, 1990; Elman, 1991) have been an essential tool in the cognitive scientist's computational toolbox for nearly 35 years. Conceptually, SRNs are typically compact neural networks that are trained to predict the next element in a sequence given the current element. A simple form of memory – a *context* layer that contains a copy of internal states from the preceding time step – provides the SRN with surprisingly robust abilities to learn and process complex recursive sequences (e.g., Elman, 1991; Cleeremans & McClelland, 1991; Cleeremans, Servan-Schreiber, & McClelland, 1989). SRNs have been applied to a wide range of perceptual, cognitive, and action domains, and were a central tool in paradigmatic shifts in cognitive theory near the end of the last century (see in particular Elman, 1996). SRNs have been and continue to be used extensively to advance theoretical understanding in many domains (for reviews, see: Elman, 1996; Plunkett & Elman, 1997; Thomas & McClelland, 2023) such as memory (e.g., Botvinick & Plaut, 2004; Botvinick & Plaut, 2006), language development (e.g., Frank, Monaghan, & Tsoukala, 2019), and in particular, psycholinguistics of language processing (e.g., Crocker & Brouwer, 2023; Christiansen & Chater, 1999a; Christiansen & Chater, 1999b). The broad scope of SRNs is why the claim that SRNs are purely feedforward (Norris, 1993, which we discuss in detail below) has such broad theoretical implications. To explain why this matters, let us consider the *feedback debates* in cognitive science.

✉ James S. Magnuson
james.magnuson@uconn.edu

1 BCBL, Basque Center on Cognition Brain and Language, Donostia-San Sebastián, Spain

2 Ikerbasque, Basque Foundation for Science, Bilbao, Spain

3 University of Connecticut, Storrs, CT, USA

4 Carnegie Mellon University, Pittsburgh, PA, USA

Published online: 13 November 2024

## Feedback debates

A central, recurring debate in the cognitive and neural sciences is whether our perceptual, cognitive, and motor capacities are autonomous (purely feedforward and/or encapsulated from cognitive influence; e.g., Norris, McQueen, & Cutler, 2000; Firestone & Scholl, 2016) or interactive (feedback systems where inputs or lower levels of representation are directly modulated by higher levels; e.g., Clark, 2013; Gilbert & Li, 2013; McClelland, Mirman, & Holt, 2006; Magnuson, Crinnion, Luthra, Gaston, & Grubb, 2024; Lupyan, 2015; Lupyan, Rahman, Boroditsky, & Clark, 2020; Proffitt, 2006; Proffitt, 2013; Schnall, 2017a; Schnall, 2017b). There are at least two levels of this debate. One is whether there exist encapsulated perceptual or cognitive modules, which may take input from other modules but are not influenced by them while carrying out their modular functions (Fodor, 1983), or whether there is continuous interaction among relatively specialized systems (Spivey, 2006). The second level, which is our concern here, is whether there is interaction *within* perceptual or cognitive systems, where prior knowledge (e.g., knowledge of the words in your language) directly influences perceptual processing (such as encoding of phonological segments).

The debate has perhaps had the greatest theoretical importance and staying power (i.e., irreconcilability) in the psycholinguistic domain of spoken word recognition. Key origins of the debate emerged with various discoveries of apparently top-down effects in spoken and visual word recognition. For example, participants are faster to identify a letter like 'A' in the context of a printed word (e.g., SCAT) than in isolation (Reicher, 1969), just as listeners are faster to detect a sound such as /k/ in the context of a spoken word (e.g., /kæat/ [CAT]) vs. a spoken pseudoword (e.g., /kʊt/, which rhymes with *soot*; Rubin, Turvey, & Van Gelder, 1976). Such *word-superiority effects* motivated theories that letters or sounds in lexical contexts receive two sources of input: a bottom-up signal, and also top-down support from lexical representations (a prime motivation for the development of interactive activation models, e.g., Rumelhart & McClelland, 1982; McClelland & Rumelhart, 1981; McClelland & Elman, 1986). Thus, as the orthographic pattern CAT or spoken word /kæt/ is experienced (even as the bottom-up encoding of the target letter or sound is progressing), any activated words send supportive feedback to their constituent elements (letters or phonemes), speeding their activation (as simulated in interactive activation models we have just cited).

There are numerous additional examples of apparent top-down effects. Ganong (1980) discovered that if participants have to identify steps from a continuum from one sound to another (e.g., /l/ to /r/) as one endpoint category or the other, the results change depending on whether one endpoint, both endpoints, or neither is a word. If the continuum is from 'ull' to 'ur' (two nonwords), we will observe an S-shaped identification pattern, with items close to the 'ull' endpoint mostly identified as containing /l/, items close to the 'ur' endpoint mostly identified as /r/, and a fairly steep shift across the middle of the continuum. If both endpoints are words (e.g., BALL to BAR), the result would be similar (though word-frequency matters; Connine, Titone, & Wang, 1993; Politzer-Ahles, Lee, & Shen, 2020). If instead one endpoint is a word (GALL) and the other a nonword (GAR), the steep part of the curve shifts away from the lexical endpoint (that is, more continuum steps are identified as the lexically consistent sound [/l/ in this example], and the shift would go in the opposite direction if the /r/ item was a word [e.g., CHAR vs. CHALL]). This lexically mediated phoneme restoration is commonly called simply 'the Ganong effect.' Similarly, if a phoneme in a word is replaced by noise, participants have great difficulty discerning whether any part of the speech was missing rather than there just being noise added to the signal (Warren, 1970). In contrast, listeners are much better at such judgments when there is no lexical context (e.g., for isolated segments; Samuel, 1981a). Generally, listeners' expectations interact with bottom-up acoustic signals to modulate speech perception (e.g., Samuel, 1981b, 1996, 1991, 1997).

Norris et al. (2000) argued that all existing demonstrations of top-down effects could be explained by an autonomous, fully feedforward process where lexical knowledge could be consulted post-perceptually. In the domain of spoken word recognition, there is a multitude of results consistent with top-down modulation of sublexical encoding in speech processing (e.g., Cibelli, Leonard, Johnson, & Chang, 2015; Getz & Toscano, 2019; Gow, Segawa, Ahlfors, & Lin, 2008; Gow & Olson, 2015; Leonard, Baud, Sjerps, & Chang, 2016; Myers & Blumstein, 2008; Noe & Fischer-Baum, 2020; Samuel, 1997; Samuel, 2001). While we disagree with the claim that autonomous models could simulate all (or even many) of these (few have actually been simulated with autonomous models), Norris et al. (2000) acknowledged that there was one paradigm that they agreed is consistent with interactive models but incompatible with autonomous models: *Lexically Mediated Compensation for Coarticulation* (LCfC; for a more recent discussion, see Norris, McQueen, & Cutler, 2016).[1]

LCfC was an innovation by Elman and McClelland (1988). Their paradigm draws on a top-down effect (Ganong) and a phonetic effect called *Compensation for Coarticulation* (CfC). In CfC (Mann & Repp, 1980; Repp & Mann,

[1] Samuel (2001) devised a paradigm that provides equally or more compelling support for interaction than LCfC: Lexically restored phonemes can potently drive selective adaptation. Proponents of autonomous models tend to dismiss this paradigm based on objections similar to those they raise about LCfC (e.g., suggesting that transitional probabilities of some degree could explain the results without appealing to feedback), which we discuss shortly.

1981), when a sound with a front place of articulation (POA) (e.g., /l/) is followed by a sound that is ambiguous between another front sound (/t/) and a back POA sound (/k/), listeners tend to identify the second sound as the one with the more back POA. Repp and Mann (1981) explained this as due to coarticulation: When a speaker must transition from a front POA to a back POA, physical constraints will make them less likely to reach the canonical back POA for the second segment. The opposite happens after a back POA segment (after /r/ with back POA, the ambiguous segment between /t/ and /k/ is more likely to be identified as /t/, which has front POA). Elman and McClelland modified the CfC paradigm by making the initial context segment also ambiguous between front and back POA. They reasoned that if that ambiguous segment could be restored based on lexical context (e.g., as /l/ given GU# [consistent with GULL], where # is ambiguous between /l/ and /r/, vs. as /r/ given CHA# [consistent with CHAR]) directly affecting the phonetic level, then the restored phoneme should drive CfC on the following ambiguous context (given a sequence like /cha#/-/#ul/, where lexical context would restore the first # as /r/ but the second would be ambiguous between TOOL and COOL). If 'restoration' is actually post-perceptual, then it should have no effect.

Elman and McClelland reported robust LCfC. However, Pitt and McQueen (1998) tested the hypothesis that LCfC could be due to sublexical regularities, such as transitional probabilities (TPs) between phonemes. They reported robust TP-mediated CfC using nonword contexts with strong TPs, and failure to observe LCfC using words with neutral transitional probabilities. Norris et al. (2000) cited this as evidence that the mediation in LCfC was sublexical and consistent with a feedforward system. They, like Pitt and McQueen, also cited a chapter by Norris (1993) as evidence that LCfC could be simulated by a feedforward system sensitive to transitional probabilities. However, that 'feedforward system' was an SRN.

This raises our key question: Are SRNs autonomous, feedforward systems, or interactive systems with feedback? In the next section, we present simple mathematical demonstrations that SRNs are not feedforward systems. Then we will return briefly to the details of Norris (1993), and the theoretical implications for SRNs as models of spoken word recognition and other aspects of human perception and cognition.

## Feedforward, feedback, and recurrent neural networks

Feedforward and recurrent networks are clearly distinguished in terms of formal definitions. For example, Prince (2023) puts it like this: "Neural networks in which the connections form an acyclic graph (i.e., a graph with no loops ... ) are referred to as feed-forward networks" [p. 35], while "recurrent neural networks [are] networks for processing sequences, in which the previous output is fed back as an additional input as we move through the sequence..." [p. 203]. That is, recurrent networks have loops, and this puts them outside the class of feedforward networks. Even one step of recurrence, as in an SRN, requires a loop. Similarly, Jurafsky and Martin (2024) put it like this: "A recurrent neural network (RNN) is any network that contains a cycle within its network connections, meaning that the value of some unit is directly, or indirectly, dependent on its own earlier outputs as an input." [ch. 8., p. 1].

To assess whether SRNs are feedforward, let us first consider the architecture of a standard *feedforward network*, as schematized in Fig. 1. Every input node has a weighted forward connection to every hidden node, and every hidden node has a weighted forward connection to every output node. Nodes sum their inputs and then apply an activation function (typically a nonlinear one, such as a sigmoid or hyperbolic tangent). Activations then feed-forward to the next layer via weighted connections. So the hidden layer input is calculated as the input vector multiplied by the input-to-hidden weight matrix. Hidden layer activations are calculated by applying an activation function (also typically nonlinear). The first processing step is when inputs are multiplied by the input-to-hidden weights to calculate hidden layer activations (Eq. 1). Then the hidden layer activations multiplied by hidden-to-output weights serve as the input to the output layer (Eq. 2), and output activations are calculated by applying activation function $g$ (which may or may not be the same as $f$) to each node's summed inputs (Eq. 3). The essential characteristic here is that the feedforward network is an *acyclic* graph: There are no cycles (loops) within the network, and so no way for information to flow in any direction except forward. In particular, the hidden layer activations depend solely upon bottom-up inputs and the input-to-hidden weights.

$$B = I \times W_{ih} \tag{1}$$

$$H = f(B) \tag{2}$$

$$O = g(H \times W_{ho}) \tag{3}$$

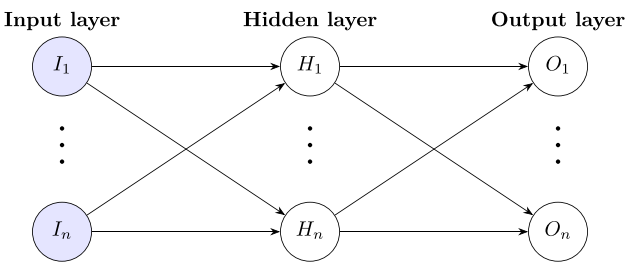


**Fig. 1** Basic feedforward network architecture. Layers are fully connected only in the forward direction (every node at the inferior level has a forward, tunable, weighted connection to every node at the superior level). *Vertical dots* stand for nodes from 2 to *n*-1, which are not depicted. Reproduced from Magnuson (2022a), with shading added

Now consider the *feedback network* depicted in Fig. 2, where we have added feedback connections from output to hidden nodes. Now the hidden units receive two sources of input: the actual bottom-up inputs, but also top-down input from the output nodes. Now the hidden unit states follow from multiplying the inputs by the input-to-hidden weight matrix *and* multiplying the output activations by the output-to-hidden weights. The hidden inputs are the sum of those two vectors, to which the activation function is applied. So while in the feedforward network the hidden units were only influenced by bottom-up inputs, now they are influenced by both bottom-up inputs and top-down feedback. This means that the activations in the hidden layer are an inextricable mixture of bottom-up and top-down signals. In contrast to the feedforward network, the feedback network graph is *cyclic*: i.e., it has loops (cycles) that allow information to flow both forward and backward.

To put this mathematically, consider Eqs. 4-7. Equation 4 describes the bottom-up input to the hidden layer (inputs multiplied by input-to-hidden weights, identical to Eq. 1). Equation 5 describes the top-down input to the hidden layer (output activations from the previous time step multiplied by output-to-hidden weights). Equation 6 states that the new hidden activations, $H$, result from summing $B$ and $T$ and then applying activation function $f$. Since $B$ and $T$ are summed, there is no way to distinguish the two sources of inputs to the hidden layer. This is the crux of objections to feedback (Norris et al., 2000): Bottom-up inputs are immediately mixed with top-down signals, making truly veridical perception impossible. Equation 7 describes the calculation of output activations. The crucial difference compared to the feedforward network is that here, the hidden layer activations depend on both bottom-up inputs ($B$) and top-down feedback ($T$).

$$B = I \times W_{ih} \tag{4}$$

$$T = O_{t-1} \times W_{oh} \tag{5}$$

$$H = f(B + T) \tag{6}$$

$$O = g(H \times W_{ho}) \tag{7}$$

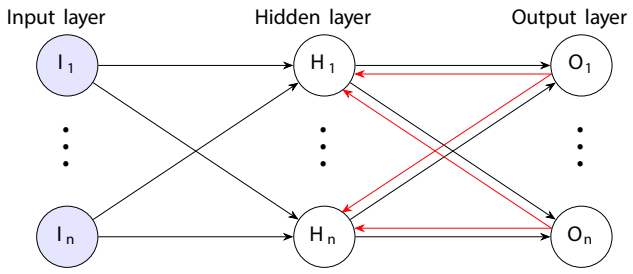


**Fig. 2** Basic feedback network architecture. We simply add connections from one layer back to nodes in a lower layer (highlighted as *red connections* here). Reproduced from Magnuson (2024a)

Now let us consider the SRN architecture, as depicted in Fig. 3. We have all the components of the feedforward network, but with the addition of *context units*. These are activated via special, 'copy-back' connections that simply duplicate the activation of corresponding hidden nodes from the previous time step to the context nodes. For example, the first context node has only one connection from the hidden layer, from the first hidden node. Context nodes are fully connected to hidden nodes in the forward direction (i.e., every context node has a tunable, weighted connection to every hidden node, including its own source node). This innovation provides the network with the ability to retain information over multiple processing steps. Depending on the pressures implied by the input–output mapping to be learned, the network can learn to encode and retain information over many time steps (by tuning the context-to-hidden weights, primarily, though the input-to-hidden weights can also participate). Elman (1990, 1991) famously innovated *next-item prediction* as a means for a network to be trained by 'self-supervised' learning, where the network attempts to predict the next item in a sequence and then uses the discrepancy between its prediction and the actual next input as the error signal for training.

Let us consider this concretely with simple equations. Equation 8 describes the part of the input to the hidden layer that is purely bottom-up: inputs multiplied by input-to-hidden weights (just as in Eq. 4). Equation 9 describes what we will provisionally identify as the top-down component of

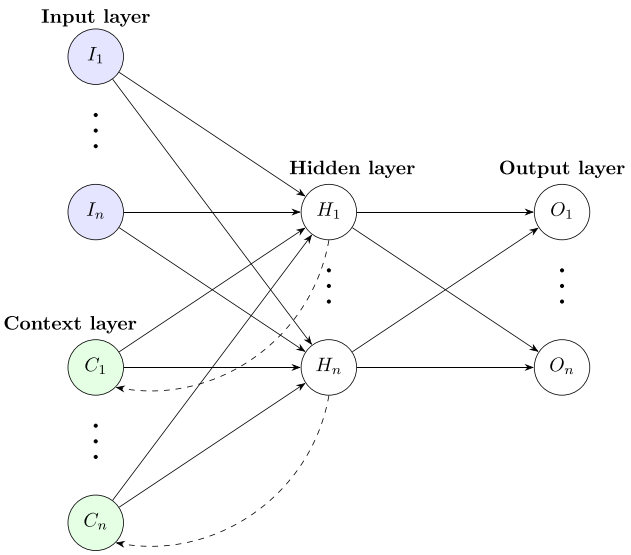


**Fig. 3** Simple Recurrent Network (SRN) architecture. Layers are fully connected in the feedforward direction (every node at the inferior level has a tunable, weighted connection to every node at the superior level). *Dashed lines* indicate 1-to-1 feedback connections (i.e., copy connections with a fixed weight of 1.0 and a one-step time delay) from each hidden unit to its corresponding context unit. *Vertical dots* stand for nodes from 2 to *n*-1, which are not depicted. Reproduced from Magnuson (2022b), with shading added

inputs to the hidden layer: the context activations multiplied by the context-to-hidden weights. Equation 10 (identical to Eq. 6) puts these together, and the new hidden activations, $H$, result from applying the nonlinear activation function, $f$, to the sum of the two input components. Again, since $B$ and $T$ are summed, there is no way for the network to distinguish bottom-up from top-down contributions to $H$.

$$B = I \times W_{ih} \tag{8}$$

$$T = C \times W_{ch} \tag{9}$$

$$H = f(B + T) \tag{10}$$

## Are SRNs autonomous or interactive?

Now we turn to the question of whether SRNs are feedforward and autonomous or have feedback and are therefore interactive. Norris (1993) asserts that a diagram like Fig. 3 facilitates understanding of processing, but falsely gives the impression that there are feedback connections in the network. He suggests that this can be appreciated if we redraw the network as in Fig. 4. Now the red connections are from each hidden node to every hidden node (including self-connections), with a time delay of one step. These are tunable, weighted connections that will be trained when the model is trained.

Norris (1993) asserts that a diagram like this "...shows that the delay connections are not really 'top-down' connections, they are just connections between the hidden units" (p. 217) and "...demonstrates that the network should properly be thought of as a bottom-up system with delay connections between the hidden units. No information is passed back down to the input units" (p. 218).

However, the error here is that the input nodes *have no function* in the network. They are in fact external, and function as placeholders or conduits for transmitting inputs to the parts of the network that *do* things (e.g., integrate inputs and transmit transformed outputs). The most prominent interactive model of human speech processing, the TRACE model (McClelland & Elman, 1986), also has input nodes that only serve to transmit external values for processing within the network. In TRACE, these activate feature nodes, which send activation forward to phoneme nodes. TRACE's feature nodes do not receive feedback from phoneme nodes by default. Phoneme nodes send activation to word nodes. When TRACE operates in interactive mode, there is feedback from word nodes to phoneme nodes. The aspect of this that, e.g., Norris et al. (2000) argued is objectionable, is that when there is word-to-phoneme feedback, the phoneme nodes no longer provide a veridical, direct mapping from the bottom-up input. That is, the bottom-up input is the product of applying input values to feature nodes, which multiply their activations by the feature-to-phoneme weights; but when feedback is 'on', phonemes also receive top-down input that is the result of multiplying phoneme activations by phoneme-to-word weights, and then multiplying resulting word activations by word-to-phoneme weights (as in Fig. 2). This is the critical level where top-down products are inextricably mixed with bottom-up products.

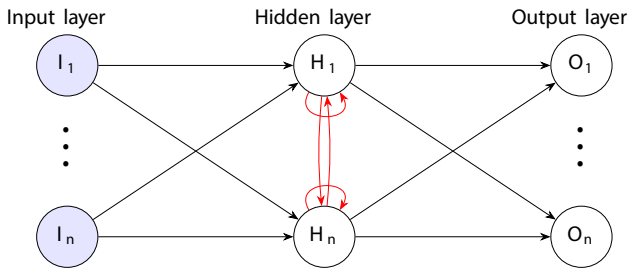


**Fig. 4** SRN architecture redepicted with hidden-to-hidden connections with a time delay (*red lines*) instead of with context nodes. Reproduced from Magnuson (2024b)

Note that the Merge model (Norris et al., 2000) also has input nodes that do not do anything except transmit stimulation to upper levels. The phoneme inputs feed-forward to a lexical layer and a phoneme decision level (which also gets input from the lexical level). The key to the Merge architecture is not protecting the input nodes, but rather performing lexical-phonemic integration outside the bottom-up pathway. In other words, the key goal is to isolate the lexical layer so it cannot influence sublexical representations, such that lexical activations are driven only by bottom-up signals and are not contaminated by mixing model-internal computations with the bottom-up pathway.

Thus, the key question is not whether there are nodes that encode the pure bottom-up inputs – TRACE has that even in interactive mode, and so does Merge. The key question is whether there is a loop in the network that mixes top-down information with bottom-up information within the bottom-up pathway (as in TRACE, where lexical feedback modulates phoneme activity, and so lexical activations are dependent on their own prior activations, but not in Merge, where lexical-phonemic integration occurs post-lexically, without affecting the lexical layer activations).

As a further illustration, consider again the SRN depicted as having lateral hidden-to-hidden time-delay connections (Fig. 4). The first point where the external input to the SRN has any impact on operations carried out by the SRN is when they arrive at the hidden nodes. However, the hidden nodes have two sources of input: the bottom-up signal multiplied by the input-to-hidden weights, *and the hidden activations from the previous time step multiplied by the hidden-to-hidden weights*.

Let us put this in mathematical form. Equation 11 states that the bottom-up input to the hidden layer, $B$, is the product of inputs and the input-to-hidden weights (note that it is

identical to Eqs. 4 and 8). Equation 12 states that the top-down (model-internal) input to the hidden layer, $T$, is the product of the hidden activations at the previous time step and the hidden-to-hidden weights (note that it is analogous to Eqs. 5 and 9, except the context states are redescribed as the previous hidden activations multiplied by the hidden-hidden weights; note also that the hidden-to-hidden weights correspond exactly to the context-to-hidden weights from Eq. 9). Equation 13 states that the new hidden activations, $H$, are the *sum* of $B$ and $T$ pushed through the activation function (and is identical to Eqs. 6 and 10). Note that $H_{t-1}$, the hidden activations at the previous time step, are the result of combining the previous inputs (at step $t-1$) with $H_{t-2}$, etc. Thus, at the initial point where input impinges on the SRN (when it is transmitted to the hidden layer), it is inextricably mixed with the results of processing previous inputs.

$$B = I \times W_{ih} \tag{11}$$

$$T = H_{t-1} \times W_{hh} \tag{12}$$

$$H = f(B + T) \tag{13}$$

Because $H_{t-1}$ is the *output* of the hidden layer at time $t-1$, the value of $H$ at time $t$ is 'dependent on its own earlier outputs as an input' (Jurafsky & Martin, 2024). Thus, an SRN is cyclic, and is therefore interactive to the degree that the first transformations it performs ($[I \times W_{ih}] + [H_{t-1} \times W_{hh}]$) modulate the bottom-up inputs via knowledge that has been acquired by training the network. Just as Eq. 6 describes a feedback network and Eq. 10 describes an SRN depicted with a context layer, Eq. 13 describes how the hidden states, $H$, of an SRN are a mixture of bottom-up and top-down information. Changing the SRN depiction from using context nodes to lateral, time-delayed connections does not change the math, nor the fact that SRNs are interactive.[2]

## Reprise: Lexically Mediated Compensation for Coarticulation (LCfC)

Now we return to the LCfC simulations conducted by Norris (1993). These simulations used an 11-feature code for phoneme inputs and a lexicon of 50 words. Inputs were presented as a sequence of feature/phoneme patterns, and the network was trained to activate the current word and the current phoneme. Feature patterns were adjusted context-sensitively at word boundaries, such that place features varied with the context of the preceding phoneme (shifted slightly back following a back POA, and slightly forward following a front POA). The network learned those segment-to-segment contingencies during training, and unsurprisingly, when a context phoneme was replaced with an ambiguous pattern that was only consistent with one word given the biphone context (e.g., the preceding phonemes could only be followed by /s/ or /ʃ/), the network exhibited both Ganong (phoneme restoration) and LCfC patterns.

Strikingly, in a related earlier chapter (Norris, 1990), Norris described the same network as implementing a feedback loop from hidden units. In 1993, however, he asserted that there is no form of feedback in an SRN because the recurrent connections are from hidden nodes to hidden nodes with a time delay of one step. As we have already discussed, the recurrent hidden layer connections in an SRN (whether described as a context loop or hidden-to-hidden connections with a time delay) form a cycle and therefore the SRN includes feedback: The hidden unit states at time $t$ are dependent on hidden unit states at time $t-1$. The hidden nodes cannot distinguish which aspects of their input are external and bottom-up (from the actual input nodes) and which are internal (from hidden nodes); the external inputs are mixed with internal information, precluding veridical input encoding.

Given that the Norris (1993) chapter continues to be cited as evidence that LCfC can be simulated without feedback, there are crucial implications of our demonstration that SRNs are models with feedback. In particular, this demonstration illustrates that positive LCfC results (as observed, for example, by Luthra et al., 2021) cannot be accounted for by a purely feedforward architecture.[3]

[2] A reviewer helpfully suggested that we address two additional, subtle details about interaction. First, an architecturally interactive/feedback network could become functionally non-interactive if, during training, it set all recurrent/feedback connections to 0 (which could theoretically happen if context information were irrelevant or misleading). Second, similar logic would apply to models with 'leaky integrator' nodes, which are another class of models where previous activation states can influence processing. We can think of each leaky integrator node having a 'context' node (a memory cell storing its previous state) and a fixed, global 'leak' parameter, which governs the relative weight of a node's previous state and its current bottom-up input. Thus, inputs to leaky integrator nodes are a mix of previous model states and bottom-up inputs. Such models are similar to SRNs but differ in that they have only self-connections (vs. full context-to-hidden connectivity in an SRN) and the leak parameter is not tunable (i.e., the context-to-node weights are not modified during training). In an otherwise non-recurrent architecture (e.g., Usher & McClelland, 2001), leaky integrator nodes would make such a network interactive and no longer purely feedforward (as long as the leak parameter is not 1, which corresponds to full leakage and therefore no influence of previous model states).

[3] Note that mixed results have been observed with the LCfC paradigm. Luthra et al. (2021) note that ~ 60% of LCfC tests have provided results consistent with lexical feedback (closer to 70% with their own results included). For example, while Samuel and Pitt (2003) and Magnuson, McMurray, Tanenhaus, and Aslin (2003a) reported positive results, McQueen, Jesse, and Norris (2009) were unable to replicate the results of Magnuson et al. (2003a) even with the original materials and observed additional failures. Luthra et al. (2021) noted that few studies had pretested items to ensure that, tested separately, context items could drive Ganong restoration and that target items were subject to CfC with unambiguous context items. If the component effects cannot be observed separately, there is no reason to expect LCfC to result when context and target items are combined. Luthra et al. observed robust, replicable LCfC when they limited their items to ones that exhibited Ganong and CfC effects separately.

We note here that this core feature of SRNs – the mixing of external inputs and model-internal states – is precisely the crucial property that proponents of purely feedforward architectures object to in networks with feedback (such as the interactive activation TRACE model; McClelland & Elman, 1986). For instance, Norris et al. (2016) argued that when models modulate bottom-up signals directly with top-down influences, they necessarily induce hallucinations (since bottom-up and top-down signals are mixed during initial processing, purely bottom-up details cannot be distinguished). Since SRNs do the same, appealing to SRNs (Norris, 1993) is actually appealing to an interactive system, not a feedforward, autonomous one (for additional discussion, see Luthra, Crinnion, Saltzman, & Magnuson, 2024).[4]

Having demonstrated that SRNs are interactive, let us return to the demonstration by Pitt and McQueen (1998) that compensation for coarticulation can be influenced by transitional probabilities. We first note that evidence for transitional probabilities modulating CfC does not itself constitute evidence against *lexically*-mediated CfC; that is, both transitional probabilities and lexical knowledge could, in theory, influence CfC (indeed, the TRACE model would predict transitional probability effects even on nonword inputs via lexical-to-phoneme feedback). Furthermore, putatively lexical influences on CfC cannot be explained by appealing to transitional probabilities: In previous work, we have shown that no single order of *n*-gram or set of *n*-grams (e.g., bigrams, trigrams) can fully explain positive observations of LCfC (Luthra et al., 2021; Magnuson, McMurray, Tanenhaus, & Aslin, 2003b). Furthermore, evidence that transitional probabilities can modulate CfC does not constitute evidence against top-down effects; for example, in the TRACE model (McClelland & Elman, 1986), TP effects emerge through top-down feedback from the lexicon.

[4] Note that Cairns, Shillcock, Chater, and Levy (1995) is occasionally cited as a demonstration that recurrent networks are feedforward and can simulate compensation effects. For example, Pitt and McQueen (1998) described it this way: "Their simulations showed that the compensation effect after an ambiguous fricative can occur in a bottom-up model with no lexical knowledge [p. 349]." However, the recurrent network used by Cairns et al. *did acquire lexical knowledge*. They trained a fully recurrent network to simultaneously activate nodes representing the previous, current, and next elements in a phoneme sequence. They describe their system as "bottom-up" in the sense that it exhibits sensitivity to lexical structure without explicit lexical representations as targets. Notably, they explicitly describe recurrent connections as feedback connections, although they also claim their results show that interactivity is not necessary to simulate LCfC. As we have demonstrated here, recurrent networks are in fact interactive in the plain sense that inputs are mixed with top-down, model-internal information. As to whether a system only trained on sublexical inputs and targets has lexical knowledge, Elman (2011) provides a deep theoretical treatment of how lexical knowledge of various sorts comes to be *embodied* in a recurrent network *emergently* without explicit, pre-specified representations (similarly, see Magnuson et al., 2020, for a case where phonetic representations emerge within a recurrent network that is never trained on phonetic targets). On this view, there is no need for a literal 'mental lexicon' with discrete entries. Instead, lexical knowledge comes to be distributed across a complex system of weights learned by the network that allows it to respond context-sensitively (e.g., lexically sensitively) to sequential inputs. Thus, Cairns et al. (1995) is actually another example of interaction in recurrent networks.

## Conclusion

In this paper, we have provided a formal demonstration that SRNs are not feedforward models, as they contain loops; further, inputs and model-internal states are inextricably mixed in SRNs, just as they are in other feedback models. This has significant implications for ongoing debates over whether perceptual and cognitive systems rely on top-down feedback, particularly in the domain of spoken word recognition. Contra previous arguments (Norris, 1993), SRNs are not feedforward. This severely reduces the previously asserted coverage of autonomous theories, based on the erroneous claim that SRNs are feedforward, autonomous systems.

**Author Note** We thank Arty Samuel, Kevin Brown, and Morten Christiansen for helpful comments. SL was supported by National Institutes of Health NRSA F32DC020625. JSM's effort was supported in part by National Science Foundation grant PAC 2043903 (PI JSM), and by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through projects PID2020-119131GB-I00 (BLIS) and PID2023-149585NB-I00 (ReNeMos).

**Author Contributions** The authors conceptualized and developed the project idea together. JSM developed the network analysis approach, created figures, and outlined the first draft. The manuscript was completed collaboratively by both authors.

**Funding** SL was supported by National Institutes of Health NRSA F32DC020625. JSM's effort was supported in part by National Science Foundation grant PAC 2043903 (PI JSM), and by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through project and PID2023-149585NB-I00 (ReNeMos).

**Availability of data and materials** Not applicable.

**Code Availability** Not applicable.

## Declarations

**Conflicts of interest/Competing interests** None.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References


Botvinick, M. M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review, 111*(2), 395–429. https://doi.org/10.1037/0033-295X.111.2.395

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113*(2), 201–233. https://doi.org/10.1037/0033-295X.113.2.201

Cairns, P., Shillcock, R., Chater, N., & Levy, J. P. (1995). Bottom-up connectionist modelling of speech. *Connectionist models of memory and language* (pp. 289–310). UCL Press Limited.

Christiansen, M. H., & Chater, N. (1999a). Connectionist natural language processing: The state of the art. *Cognitive Science, 23*(4), 417–437. https://doi.org/10.1207/s15516709cog2304_2

Christiansen, M. H., & Chater, N. (1999b). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23*(2), 157–205. https://doi.org/10.1207/s15516709cog2302_2

Cibelli, E. S., Leonard, M. K., Johnson, K., & Chang, E. F. (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. *Brain and Language, 147*, 66–75. https://doi.org/10.1016/j.bandl.2015.05.005

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120*(3), 235–253. https://doi.org/10.1037/0096-3445.120.3.235

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation, 1*(3), 372–381. https://doi.org/10.1162/neco.1989.1.3.372

Connine, C. M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 19*(1), 81–94. https://doi.org/10.1037/0278-7393.19.1.81

Crocker, M. W., & Brouwer, H. (2023). Computational psycholinguistics. *The cambridge handbook of computational cognitive sciences*. Cambridge University Press. https://doi.org/10.1017/9781108755610.032

Elman, J. L. (1996). *Rethinking innateness: A connectionist perspective on development*. MIT Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*(2), 195–225. https://doi.org/10.1007/BF00114844

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The mental lexicon, 6*(1), 1–33. https://doi.org/10.1075/ml.6.1.01elm

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*(2), 143–165. https://doi.org/10.1016/0749-596X(88)90071-X

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*, e229. https://doi.org/10.1017/S0140525X15000965

Fodor, J. A. (1983). *The modularity of mind*. The MIT Press.

Frank, S. L., Monaghan, P., & Tsoukala, C. (2019). Neural network models of language acquisition and processing. https://doi.org/10.7551/mitpress/10841.003.0026

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

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science, 30*(6), 830–841. https://doi.org/10.1177/0956797619841813

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience, 14*(5), 350–363. https://doi.org/10.1038/nrn3476

Gow, D. W., & Olson, B. B. (2015). Lexical mediation of phonotactic frequency effects on spoken word recognition: A granger causality analysis of MRI-constrained MEG/EEG data. *Journal of Memory and Language, 82*, 41–55. https://doi.org/10.1016/j.jml.2015.03.004

Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F.-H. (2008). Lexical influences on speech perception: A granger causality analysis of MEG and EEG source estimates. *NeuroImage, 43*(3), 614–623. https://doi.org/10.1016/j.neuroimage.2008.07.027

Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing (3rd edition draft)*. https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications, 7*(1), 13619. https://doi.org/10.1038/ncomms13619

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology, 6*(4), 547–569. https://doi.org/10.1007/s13164-015-0253-4

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences, 24*(11), 930–944. https://doi.org/10.1016/j.tics.2020.08.005

Luthra, S., Crinnion, A. M., Saltzman, D., & Magnuson, J. S. (2024). Do they know it's christmash? lexical knowledge directly impacts speech perception. *Cognitive Science, 48*(5), e13449. https://doi.org/10.1111/cogs.13449

Luthra, S., Peraza-Santiago, G., Beeson, K., Saltzman, D., Crinnion, A. M., & Magnuson, J. S. (2021). Robust lexically mediated compensation for coarticulation: Christmash time is here again. *Cognitive Science, 45*(4), e12962. https://doi.org/10.1111/cogs.12962

Magnuson, J. S. (2022a). Feedforward network in tikz. https://doi.org/10.6084/m9.figshare.20165456.v1

Magnuson, J. S. (2022b). SRN in tikz. https://doi.org/10.6084/m9.figshare.20165324.v1

Magnuson, J. S. (2024a). *FBN: Tikz feedback network* [Figshare]. https://doi.org/10.6084/m9.figshare.25333783.v1

Magnuson, J. S. (2024b). *SRN: Tikz diagram with lateral time-delay connections* [Figshare]. https://doi.org/10.6084/m9.figshare.25333765.v1

Magnuson, J. S., Crinnion, A. M., Luthra, S., Gaston, P., & Grubb, S. (2024). Contra assertions, feedback improves word recognition: How feedback and lateral inhibition sharpen signals over noise. *Cognition, 242*, 105661. https://doi.org/10.1016/j.cognition.2023.105661

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003a). Lexical effects on compensation for coarticulation: The ghost of christmash past. *Cognitive Science, 27*(2), 285–298. https://doi.org/10.1207/s15516709cog2702_6

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003b). Lexical effects on compensation for coarticulation: The ghost of christmash past. *Cognitive Science, 27*(2), 285–298. https://doi.org/10.1016/S0364-0213(03)00004-1

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A minimal neural network

model of incremental human speech recognition. *Cognitive Science, 44*(4), e12823. https://doi.org/10.1111/cogs.12823
Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the []-[s] distinction. *Perception & Psychophysics, 28*(3), 213–228. https://doi.org/10.3758/BF03204377
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. Psychological Review, 88(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0
McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(8), 363–369. https://doi.org/10.1016/j.tics.2006.06.007
McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical–prelexical feedback during speech perception or: Is it time to stop playing those christmas tapes? *Journal of Memory and Language, 61*(1), 1–18. https://doi.org/10.1016/j.jml.2009.03.002
Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: An fMRI investigation. *Cerebral Cortex, 18*(2), 278–288. https://doi.org/10.1093/cercor/bhm053
Noe, C., & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition, 197*, 104162. https://doi.org/10.1016/j.cognition.2019.104162
Norris, D. (1990). A dynamic-net model of human speech recognition. *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 87–104). The MIT Press.
Norris, D. (1993). Bottom-up connectionist models of 'interaction'. *Cognitive models of speech processing: The second sperlonga meeting* (pp. 211–234). Lawrence Erlbaum Associates Publishers.
Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*(3), 299–325. https://doi.org/10.1017/S0140525X00003241
Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18. https://doi.org/10.1080/23273798.2015.1081703
Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*(3), 347–370. https://doi.org/10.1006/jmla.1998.2571
Plunkett, K., & Elman, J. (1997). *Exercises in rethinking innateness: A handbook for connectionist simulations*. MIT Press.
Politzer-Ahles, S., Lee, K. K., & Shen, L. (2020). Ganong effects for frequency may not be robust. *The Journal of the Acoustical Society of America, 147*(1), EL37. https://doi.org/10.1121/10.0000562
Prince, S. J. D. (2023). *Understanding deep learning*. The MIT Press.
Proffitt, D. R. (2006). Embodied perception and the economy of action. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 1*(2), 110–122. https://doi.org/10.1111/j.1745-6916.2006.00008.x
Proffitt, D. R. (2013). An embodied approach to perception: By what units are visual perceptions scaled? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 8*(4), 474–483. https://doi.org/10.1177/1745691613489837
Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*(2), 275–280. https://doi.org/10.1037/h0027768
Repp, B. H., & Mann, V. A. (1981). Perceptual assessment of fricative–stop coarticulation. *The Journal of the Acoustical Society of America, 69*(4), 1154–1163. https://doi.org/10.1121/1.385695
Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics, 19*(5), 394–398. https://doi.org/10.3758/BF03199398
Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*(1), 60–94. https://doi.org/10.1037/0033-295X.89.1.60
Samuel, A. G. (1981a). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*(4), 474–494. https://doi.org/10.1037/0096-3445.110.4.474
Samuel, A. G. (1981b). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1124–1131. https://doi.org/10.1037/0096-1523.7.5.1124
Samuel, A. G. (1991). A further examination of attentional effects in the phonemic restoration illusion. *The Quarterly Journal of Experimental Psychology Section A, 43*(3), 679–699. https://doi.org/10.1080/14640749108400992
Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*(1), 28–51. https://doi.org/10.1037/0096-3445.125.1.28
Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*(2), 97–127. https://doi.org/10.1006/cogp.1997.0646
Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science, 12*(4), 348–351. https://doi.org/10.1111/1467-9280.00364
Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48*(2), 416–434. https://doi.org/10.1016/S0749-596X(02)00514-4
Schnall, S. (2017a). No magic bullet in sight: A reply to firestone and scholl (2017) and durgin (2017). *Perspectives on Psychological Science, 12*(2), 347–349. https://doi.org/10.1177/1745691617691948
Schnall, S. (2017b). Social and contextual constraints on embodied perception. *Perspectives on Psychological Science, 12*(2), 325–340. https://doi.org/10.1177/1745691616660199
Spivey, M. (2006). *The continuity of mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195170788.001.0001
Thomas, M. S. C., & McClelland, J. L. (2023). *Connectionist models of cognition*. Cambridge University Press. https://doi.org/10.1017/9781108755610.005
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(3917), 392–393. https://doi.org/10.1126/science.167.3917.392