# The Lichtheim-memory model: A computational model of language comprehension, production, and verbal working memory

by

Steven C. Schwering

A dissertation submitted in the partial fulfillment of the requirement for the degree of

Doctor of Philosophy (Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 05/24/2023

The dissertation is approved by the following members of the final oral committee:

Maryellen C. MacDonald, Professor, Psychology

Joseph Austerweil, Associate Professor, Psychology

Timothy Rogers, Professor, Psychology

Haley Vlach, Associate Professor, Educational Psychology

Table of Contents

**Acknowledgments**

There are many people who contributed to this work, either directly or indirectly, to make my doctoral research possible. My acknowledgments cannot possibly touch on all their contributions, and there are many who helped me along the way who are not mentioned by name. I hope this dissertation can serve as a testament to their mentorship, friendship, and love. This work would not have been possible without their support.

Thank you to my advisor, Dr. Maryellen MacDonald for providing me the space and time to pursue my academic interests. While my projects did not always work out, Maryellen's guidance and encouragement always inspired me to continue my work even when grappling with both the memory literature and the language literature became overwhelming. Forward.

Thank you to my committee members, Dr. Joe Austerweil, Dr. Tim Rogers, Dr. Haley Vlach, and Dr. Jenny Saffran, who provided me guidance on my dissertation project and on my other projects throughout graduate school. My approach to science is better for their efforts. Also, thank you to my undergraduate advisors who encouraged me to apply to graduate school and gave me my first taste of research: Dr. Tamara Bireta, Dr. Andrew Leynes, and Dr. Jarrett Crawford.

Thank you to all my friends in graduate school who gave me perspective on important things in life while also giving me important feedback on my best and worst ideas. I am especially grateful for my friends from the Language and Cognitive Neuroscience Lab: Elise Hopman, Matt Cooper-Borkenhagen, Mark Koranda, Arella Gussow, Matt Borman, Cass Jacobs, and Misty Kabasa. My friends in my cohort and Dungeons and Dragons groups gave me many laughs, and we shared good food: David Menendez, Jake Richie, Nadia Doutcheva, AJ Peters, Martin Zettersten, and more.

Thank you to my undergraduate research assistants and undergraduate students who completed my many experiments. I am proud of your accomplishments, and I hope your successes far exceed mine.

A big thank you to my family. Madison became a home during graduate school, but I will always look forward to cooking and eating good food with you. Thank you to my mother for giving me the drive to succeed at difficult work and thank you to my father for giving me the frame of mind that helped me to enjoy life during difficult times. Thank you to my brother for giving me someone to look up to when I first thought of becoming a scientist.

And most importantly, I owe much of my thanks to Courtney Johnson. She saw in me a scientist when I could not. Words cannot convey how fortunate I am to have someone so thoughtful, intelligent, and caring in my life.

**Abstract**

Language comprehension and production is frequently thought to rely on a distinct verbal working memory (VWM) capacity. This perspective relies upon the notion that VWM is a wholly separate domain, supported by mechanisms distinct from language processing. In this work, I consider a language emergent alternative: that VWM is instead supported by language comprehension and production. I describe the Lichtheim-memory model, a neural network trained to comprehend, produce, and repeat both words and sentences. Then, I test the extent to which the model can perform serial recall of nonwords, noun lists, sentences, and sentence-like lists, comparing the model's performance to human memory benchmarks. The Lichtheim-memory model successfully captures general patterns of human performance solely through its language processing abilities, providing a mechanism for language emergent VWM.

## Introduction

Language use and serial recall share many striking similarities: both require encoding, maintaining, and producing words in the correct order. Proceduralist approaches to verbal working memory (VWM) have long recognized these similarities (Crowder, 1993), hypothesizing the procedures that support behavior like language comprehension and production may also maintain temporary memories. Empirical evidence supports this hypothesis, with many researchers now proposing language emergent theories of VWM in which procedures of comprehension and production encode, maintain, and reproduce temporary memory (MacDonald, 2016; Majerus, 2013; Schwering & MacDonald, 2020). However, the computational principles supporting these procedures remain underspecified (Norris, 2017). In this dissertation, I aim to specify these computations in a computational model of both language use and VWM. Specifically, I adopt principles of the rich language emergent theory of VWM (Schwering & MacDonald, 2020), in which a web of word representations, order representations, and word-order interactions are hypothesized to affect both language use and memory. I describe the Lichtheim-memory model of sentence and list repetition, comprehension, production, and serial recall. Further, I conduct several tests of the model to probe its ability to employ word representations, order representations, and word-order interactions in serial recall.

Much of this work concerns the definition of VWM and its characterization in different models. VWM is typically characterized as a form of temporary storage of information for use later, or as a resource that can store and manipulate information for other tasks (Cowan, 2008; 2017). VWM is important because it is a core cognitive construct, describing the ability of people to hold information in mind despite engaging in concurrent processing of other information. This ability means the maintained memory must be resilient in the face of decay

over time or in the face of interference from other information. Researchers characterize the mechanisms and limitations on VWM in many different ways (e.g. Portrat et al., 2005), leading to vastly different application in real world settings, like in treatment of language impairments (e.g. Wright & Shisler, 2008). Given its centrality to many aspects of cognition, as well as existing debate over its nature, better understanding the mechanisms supporting VWM is an important objective for psychological science.

**How Does the Language System Interact with VWM?**

Language comprehension and production are often thought to be constrained by VWM. According to some researchers, VWM supplies the processing capacity for the language system by binding referents in long-distance dependencies, both in the context of comprehension (Tan, Martin, & Van Dyke, 2017; Van Dyke & Johns, 2012) and production (Freedman et al., 2004; Martin & Freedman, 2001; Slevc & Martin, 2016). VWM also affects learning. Children with higher VWM capacity learn words faster than children with lower VWM capacity (Gathercole & Baddeley, 1989) by allowing language learners to maintain sequences of phonemes in a passive storage system and encode frequently encountered sequences into the mental lexicon (Baddeley, Gathercole, & Papagno, 1998; Page & Norris, 2009). Further still, VWM training is employed as a tool to aid aphasic patients recovering from traumatic brain injury (Majerus, 2018; Nikravesh et al., 2021) on the basis that deficits in language processing are rooted in deficits of VWM (Freedman et al., 2004; Martin & Freedman, 2001; Slevc & Martin, 2016). Given these many interactions between VWM and language, understanding the specific way in which VWM and language processes interact is an important step in understanding language comprehension and production.

Modular perspectives have dominated the intersection of memory and language research (Adams, Nguyen, & Cowan, 2018; Baddeley, 2017). In the modular approach, VWM is described as a distinct cognitive module, employed by the language system to maintain partially processed linguistic representations. This perspective coexists amicably with buffer theories of VWM (Baddeley & Hitch, 1974; Norris, 2017). In buffer theories, VWM is characterized as a blank slate continually overwritten by immediate experience and separated from long-term memory (LTM). The transience of VWM is intentional. According to proponents, memory buffers are necessary to represent random or novel memory lists that are unlike natural language and therefore not present in LTM (Allen et al., 2009; Norris, 2017). Therefore, VWM can encode novel words and utterances, assisting in the processing of long-distance dependencies, word learning, and recovery from aphasia.

Perhaps one of the most useful ways to specify the relationship between VWM and language is through a computational model that makes the mechanisms supporting each domain explicit. Buffer theories may be instantiated through many different computational models. For example, primacy (Page & Norris, 1998) and start-end models (Henson, 1998) characterize memory encoding through oscillator strength, and memory maintenance in a domain-independent store. Page and Norris (2009) build on primacy models to account for word learning via the Hebb repetition effect, the finding that repeated exposure to a regular sequence improves memory. Despite these successes, ties to language remain limited. Much remains to be done to computationally specify how VWM affects comprehension and production beyond word representations like phonology.

Furthermore, there are reasons to believe core assumptions of buffer theories may not accurately reflect the relationship between VWM and language. Behavioral and neuroimaging

research suggest language LTM and brain regions supporting language use also support VWM, undermining the separation between language and memory domains (see Schwering & MacDonald, 2020 for review).

Word representations support VWM. Real and regular words are recalled more often than non-words and irregular words (Hulme et al., 1995; Roodenrys, Hulme, & Brown, 1993). Similarly, frequent words (Hulme et al., 1997; Poirier & Saint-Aubin, 1996), contextually diverse words (Johns, 2021), and concrete words (Walker & Hulme, 1999) are recalled better than infrequent words, less contextually diverse words, and abstract words. Parallel results are observed in the psycholinguistic literature. Frequent (Grainger, 1990), contextually diverse (Adelman, Brown, & Quesada, 2006), and concrete words (Schwanenflugel, 1991; Schwanenflugel, Harnishfeger, & Stowe, 1988) are more easily accessed in lexical decision tasks than less frequent, less contextually diverse, and less concrete words.

Larger linguistic units also support VWM. Participants are more likely to recall common multiword phrases than random lists of words (Jacobs et al., 2016; Jacobs et al., 2017), and lists that are more sentence-like are recalled more accurately than lists that are less sentence-like (Allen, Hitch, & Baddeley, 2018; Baddeley, Hitch, & Allen, 2009; Lombardi & Potter, 1992; Potter & Lombardi, 1990). Again, parallels are observed in language use. Participants are faster to read common multiword phrases than novel sentences (Arnon & Snider, 2010), and predictable sentences are processed faster than sentences that are less predictable (Smith & Levy, 2013).

These results suggest that VWM may not just impact language, but that language may also impact VWM. Buffer theories have proposed several theoretical patches to account for this behavioral evidence. For example, redintegration repairs degraded memory traces using support

from linguistic LTM (e.g. Hulme et al. 1997). However, with an increasing number of temporary buffers for phonological, semantic, and other linguistic representations (e.g. Martin et al., 1994) buffer theories are becoming increasingly complicated; specifying the computations underpinning buffer theories, linguistic LTM, and their interactions remains a distant future direction. In response, some researchers have sought to unite VWM and language under a common framework to create a more parsimonious account of VWM and language.

**Language Emergent Theories of Verbal Working Memory**

Language emergent theories of VWM reject the distinction between VWM and language processes. Rather, inspired by proceduralist approaches to VWM (Crowder, 1993), language emergent theories characterize memory through language comprehension and production procedures. Under this view, language comprehension systems encode temporary memories and language production systems maintain and enact the plan to reproduce comprehended memoranda (Acheson & MacDonald, 2009; MacDonald, 2016; Majerus, 2013; Schwering & MacDonald, 2020). This language emergent approach suggests a vastly different computational architecture may support VWM and language-memory interactions than the one proposed by buffer theories of VWM.

Specific language emergent theories characterize language differently or specify limits on the extent to which the language system supports VWM, each providing a different picture of how VWM may emerge from language use. Limited emergent theories of VWM argue that word comprehension and production processes govern memory for words in VWM tasks, but the ability to recall the order of those words is governed by a separate memory buffer (Majerus, 2013). The key notion behind this perspective is the language comprehension and production systems can maintain memory for familiar words, but not unfamiliar, random memory lists. As a

result, memory for the order of words must be governed by a separate VWM system that processes novel sequences. These arguments reflect many of the classic perspectives in the memory literature, which suggest a VWM buffer must exist to handle processing of novel stimuli (Hitch, Hurlstone, & Hartley, 2022; Norris, 2017).

The rich emergent theory of VWM discards the notion that memory for words and memory for order are governed by separate systems. Rather, this theory argues that the language system learns to employ integrated representations of words and their context to best complete the tasks of language comprehension and production, and these same representations underlie performance in VWM tasks (Schwering & MacDonald, 2020). Rich emergent theory borrows heavily from constraint satisfaction theories of language processing (Seidenberg & MacDonald, 1999; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993). According to constraint-based perspectives, representations underlying performance in language tasks integrate classic linguistic features described by phonology, morphology, semantics, syntax, and more (MacDonald, 1994; MacDonald et al., 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, & Hann, 2000). This perspective suggests that the separation of word representations and word order representations in limited emergent perspectives may be incorrect. If word and word order representations are integrated in linguistic LTM, and linguistic LTM informs VWM, then word and word order must be integrated in VWM.

I mention each of these phenomena because they constrain the mechanisms that can support language comprehension, language production, and VWM. Being grounded in the procedures of language comprehension and production (Crowder, 1993), emergent computational models of VWM must account for memory phenomena through language comprehension and production mechanisms. This has proven to be challenging. The rich

emergent perspective has been stymied by a dearth of computational models that make its theoretical commitments explicit. This is clearly illustrated in Norris (2017) which adopts a grim perspective of emergent models, writing, “there are no computational models [of VWM] based on activation of LTM. Before activation-based models can be taken seriously they need to be clearly formulated and… shown to be competitive with existing computational models” (p. 999). In the next sections, I seek to address the complaint levied in Norris (2017). First, I identify the behavioral phenomena that must be captured by a rich emergent model of VWM. Emphasis is paid to phenomena critical to supporting the rich emergent theory over alternatives. Then, I outline a computational model consistent with the rich emergent theory that can account for this behavior to better bridge the language and memory domains.

**Behavioral Phenomena: Targets of a Rich Emergent Computational Model**

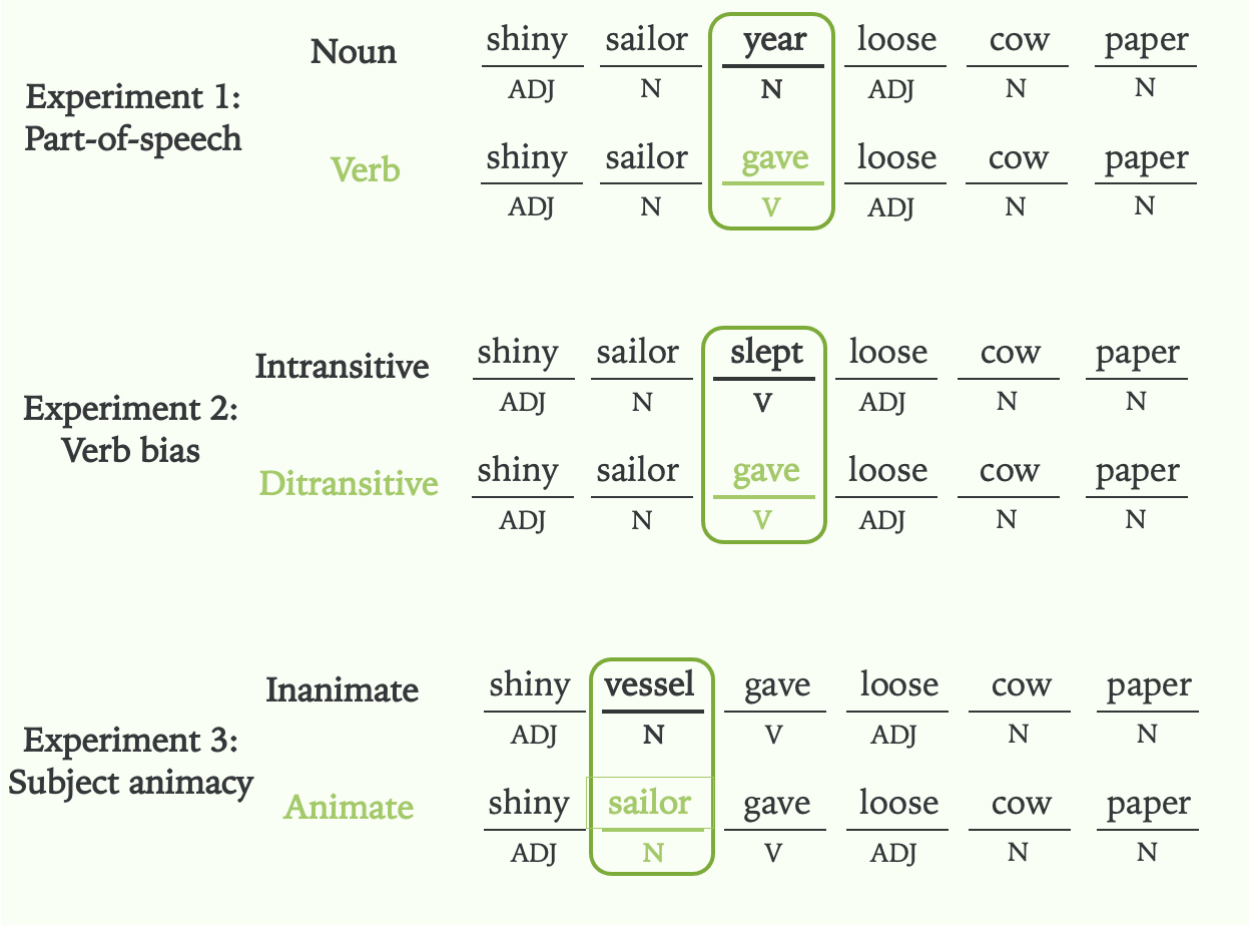
Memory for sentence-like lists provide a particularly rich testbed to examine how integrated representations affect VWM. Sentence-like lists tend to be recalled better than randomly generated memory lists (Allen, Hitch, & Baddeley, 2018; Baddeley, Allen, & Hitch, 2009). List-wide properties moderate this effect, such as experimenter-defined meaningfulness of the sentence (Jones & Farrell, 2018). Minute grammatical regularities also affect memory. Lists of adjective-noun pairs in canonical order (e.g. *hostile window*) tend to be recalled better than pairs in the reverse order (e.g. *window hostile*; Perham, Marsh, & Jones, 2009), and this effect is enhanced when morphosyntactic properties reinforce the grammatical regularity (Schweppe et al., 2022). These findings are consistent with the *rich emergent* perspective: properties of words (i.e. the part-of-speech of a word) interact with their context (i.e. the ordering of the words) to inform VWM.

Lexico-syntactic constraints are a prime target for observing word-order interactions in both language and memory. Lexico-syntactic constraints are word properties that influence comprehension and production of the syntactic structure of sentences. For example, word properties like the typical part-of-speech in which a word occurs, strongly constrain comprehension and production of sentence structure. Verb biases, the statistical bias for verbs to appear in specific sentence structures, affect reading times of ambiguous sentences (Trueswell et al., 1993). Animacy, a semantic property of nouns, biases the event structure of sentences and facilitates processing of sentences with animate subjects over inanimate subjects (Szewczyk & Schriefers, 2011). Critically, the impact of lexico-syntactic constraints depends upon the interaction of specific lexical features (e.g. part-of-speech, verb biases, animacy) and the syntactic context in which they are embedded; mutual constraints between lexical properties and syntactic structure impact language processing in ways that cannot be predicted by lexical properties or syntactic structure alone (MacDonald et al., 1992; MacDonald et al., 1994; Seidenberg & MacDonald, 1994). If lexico-syntactic constraints are a critical component of language comprehension and production, and the language system supports VWM, then integrated word and order representations encoded in lexico-syntactic constraints should support VWM.

This exact prediction was tested across 3 serial recall experiments described in Schwering et al. (under review). In these experiments, participants recalled simplified ditransitive sentences in which a single word varied in part-of-speech, verb bias, or animacy. Example memory lists are illustrated in *Figure 1*. Across all conditions, participants were more likely to recall ditransitive sentence-like lists when lexico-syntactic constraints of the manipulated word supported the ditransitive sentence structure than when it did not, even when controlling for other lexical features like length, frequency, and contextual diversity. These results are illustrated in *Figure 2*.

**Figure 1**

*Example memory lists manipulating lexico-syntactic constraints*

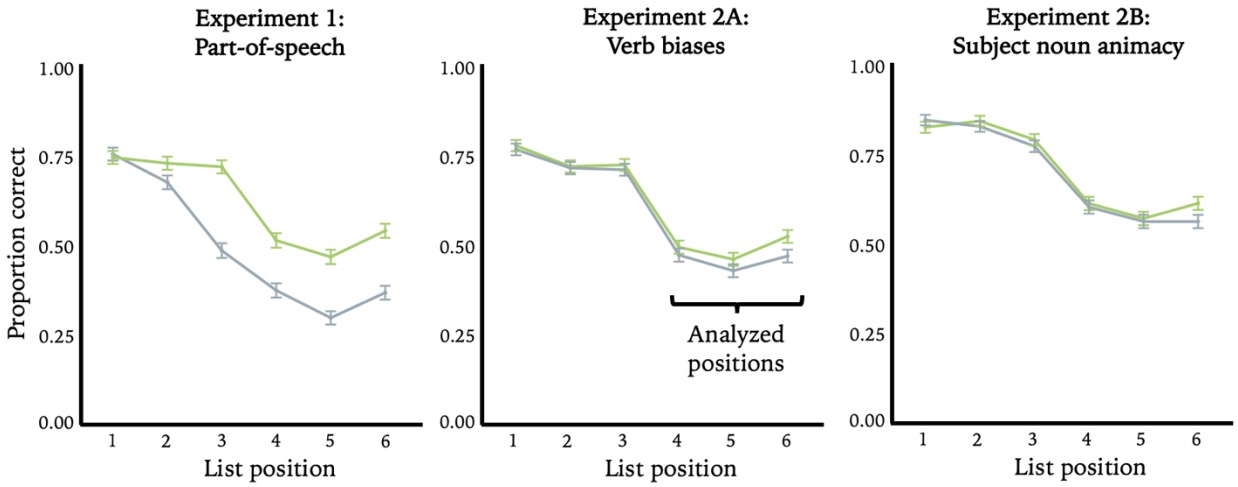



*Note.* Each row of words (e.g. *shiny sailor year loose cow paper*) represents a sample memory list. One lexico-syntactic feature was manipulated between conditions. The manipulated word in each experiment is illustrated in the green box, with the condition bolded to the left of the list. All conditions supported by the lexico-syntactic constraint (i.e. verb, ditransitive verb, animate subject) are illustrated in green. Alternatives not supported by the lexico-syntactic constraint (i.e. noun, intransitive verb, inanimate noun) are illustrated in black. All other words outside of the manipulated word were held constant between conditions. If lexico-syntactic constraints support

memory, then lists supported by the green lexico-syntactic constraint should be recalled better in position than lists supported by the black lexico-syntactic constraint.

**Figure 2**

*Impact of lexico-syntactic constraints on serial recall*




*Note.* Green bars represent recall in position for sentence-like lists supported by the lexico-syntactic constraint. Gray bars represent recall in position for sentence-like lists not supported by the lexico-syntactic constraint. All differences between conditions were significant, meaning participants were more likely to recall sentence-like lists supported by the lexico-syntactic constraints than alternatives.

These data demonstrate VWM capacity is sensitive to word-order interactions, because the lexico-syntactic constraints that support memory are a function of both word properties (i.e. part-of-speech, verb bias, noun animacy) and the sentence context in which they are embedded. These results are taken as support for the rich emergent theory of VWM, which argues the language system supports memory for both words and the orders in which they occur. This perspective

contrasts sharply against both buffer theories of VWM (Norris, 2017), which argues a separate VWM capacity maintains memory for sentence-like lists, and limited emergent theories of VWM (Majerus, 2013), which argues the language system can support word memory but not memory for order or novel lists.

Multiple computational architectures may account for the basic sentence superiority effect and for effects of part-of-speech on sentence recall (i.e. Experiment 1 in Schwering et al., under review). For example, Jones & Farrell (2018) account for the effect of part-of-speech on the sentence superiority effect through redintegration, consistent with buffer and limited emergent theories of VWM. In the model, memory is enhanced through a redintegration mechanism in proportion to part-of-speech sequence frequency in natural language. LTM intervened to rebuild degraded memory traces stored in a buffer. Accounting for the effects of verb biases and subject noun animacy on VWM remain challenging for computational models.

Theoretically, other computational models consistent with the rich emergent theory of VWM, could be developed to account for these patterns. In these approaches, linguistic LTM encoding lexico-syntactic constraints should support memory directly (Schwering & MacDonald, 2020). Lexico-syntactic constraints are a common element of computational models of language processing (e.g. Joanisse & Seidenberg, 2003; Monaghan & Woollams, 2017; Seidenberg & McClelland, 1989). However, no rich language emergent model adopting similar principles has been applied to VWM. Defining a computational architecture that can account for the impact of lexico-syntactic constraints on VWM is an important next step for rich emergent theories of VWM.

Of course, lexico-syntactic constraints are not the only memory phenomena important to the rich language emergent theory of VWM. Any computational model of VWM must account for general memory benchmarks, like those outlined in Oberauer et al. (2018). Two additional

phenomena are of particular interest given their importance in highlighting interactions between VWM and linguistic LTM for words and orders. Real or regular words tend to be recalled better than non-words (Hulme et al., 1995), and sentences tend to be recalled better than the same list of scrambled words (Allen et al., 2018). If linguistic LTM for words and orders impact VWM, then word superiority and sentence superiority should be observed in a rich emergent model of VWM alongside sensitivity to supportive lexico-syntactic constraints.

**Characterizing Integrated Word-order Representations in Memory Models**

What computational principles should guide a rich emergent theory of VWM? Despite Norris (2017)'s claim that "there are no computational models [of VWM] based on activation of LTM" (p. 999), many memory models have already made strides to instantiate emergent theory. While these models fall short of instantiating a rich emergent theory, they nevertheless provide important insight into the kinds of computations that may be relevant.

Before describing each of these emergent models, it is worth noting why these models can be described as emergent. Many of the models discussed below are neural networks. In these neural networks, activity is transformed by a series of weights in the service of some task. These weights instantiate the models' LTM and are either learned by networks through trial and error or are set by a researcher to achieve the models' objectives. Model activity, transformed by the models' weights (i.e. their LTM), represent the models' VWM. Characterizing the models as a language emergent requires model weights to transform activity in such as a way as to perform language tasks. Neural networks have frequently been employed in the language literature to understand how language experience may inform linguistic LTM and language comprehension and production mechanisms (e.g. Joanisse & Seidenberg, 2003; Monaghan & Woollams, 2017; Seidenberg & McClelland, 1989). While neural network models may not often be recognized by

memory researchers as instantiating an implicit, temporary memory, language researchers have long considered this a valid comparison (e.g. MacDonald & Christiansen, 2002; Martin et al., 1996).

Most emergent models of VWM have adopted principles of lexical selection or have otherwise focused on word representations. For example, Martin et al. (1996) developed a two-stage model of lexical selection adopting principles of spreading activation between phonological, lexical, and semantic representations. Lesioning weights in the network simulated aphasic patients' comorbid VWM and language deficits, while an intact network simulated healthy participant performance. More recent models, like the semantic network applied to serial recall in Kowialiewski et al. (2021), also adopt word representations in support of memory. Kowialiewski et al. (2021) characterized a language emergent model through activation in a semantic network, treating memory as spreading activation within the semantic network and recall as iterative selection and inhibition of the most active semantic representations. While these models neatly fit within limited emergent perspectives, which argue VWM is supported by word representations, they fail to capture critical interactions between lexical and syntactic representations that are central to rich emergent perspectives. As a result, word-oriented models exhibit a poor capacity to capture order memory (Kowialiewski et al. 2021), limiting their application to VWM research and the larger language literature.

Other models in the emergent vein integrate item (i.e. word) representations and order representations, though not in a way comparable to natural language processing. For example, Botvinick and Plaut (2006) trained a simple recurrent neural network to perform serial recall, successfully capturing general serial recall phenomenon like serial positions curves, with some ability to generalize to novel sequences. The model learned to perform serial recall through

experience, by learning to encode conjoined item and order representations. Similarly, Gupta and Tisdale (2009) applied the simple recurrent architecture to word learning, forcing the model to learn phonotactic regularities of its training set. However, these models fail to capture important lexico-syntactic constraints that impact language use and memory, like part-of-speech, verb biases, and animacy (Schwering et al., under review.). Botvinick and Plaut (2006) trained networks on arbitrary patterns or a simplified grammar lacking lexico-syntactic constraints, and Gupta and Tisdale (2009) modeled phonotactics. While these models could, in principle, employ representations akin to lexico-syntactic constraints, their ability to capture lexcio-syntactic constraints like part-of-speech, verb biases, and animacy remains untested.

The last class of emergent models fail to fully embrace principles of emergent theory by attaching a memory buffer to a language-inspired processing architecture. For example, Hartley, et al. (2016) generated a model of phonological working memory based on principles of competitive cueing, in which memory for an acoustic signal is encoded through parallel oscillators inspired by neural oscillations in language processing. While temporary memory is encoded through a language-inspired mechanism, memory itself is not stored in the language processing system; the memory is passed to a separate memory buffer. This general approach is also modeled in Page and Norris (2009), which attaches a primacy gradient buffer (Page & Norris, 1998) to phonotactic LTM. This class of models captures the perspective that VWM is linked with but functionally separate from language processing and is thus incompatible with rich emergent theory.

Four primary principles of a rich emergent computational model may be distilled from these models. First, as demonstrated by both Martin et al. (1996) and Botvinick and Plaut (2006), a rich emergent model must characterize VWM as the activated form of LTM. This is a general

principle of language emergent models and directly contradicts buffer theories. Second, as demonstrated by Martin et al. (1996), the LTM of a rich emergent model must comprise language comprehension and production procedures. Specific ties to language processes like lexical selection naturally accommodate this principle. Third, as demonstrated by Botvinick and Plaut (2006), the LTM of a rich emergent model must integrate word and order representations. In conjunction with the previous principle, this suggests that comprehension and production of words and sentences may serve as a prime target for a rich emergent model. Fourth, a rich emergent model must instantiate VWM without the use of buffers.

Artificial neural networks provide a natural way to incorporate all these principles. Under the right demands, neural networks learn integrated word-order representations consistent with constraint-based approaches in language comprehension and production (e.g. Joanisse & Seidenberg, 2003; Monaghan & Woollams, 2017; Seidenberg & McClelland, 1989). Neural networks learn these representations through experience and encode these representations in their LTM weights. Given the right language tasks, a neural network could serve as the foundation of a rich emergent computational model of VWM, specifying the computations through which lexico-syntactic constraints support VWM as well as the computations that support more general memory phenomenon like word superiority and sentence superiority effects.

## Modeling: Can a language model account for behavior in serial recall tasks?

Generating a model of the rich emergent perspective is paramount to contrasting rich emergent theory against both buffer theories and limited emergent theory: computational models make theoretical claims explicit and allow the development of novel predictions that may otherwise be difficult to explain or understand without the aid of the model. In this project, I develop a computational model of both VWM and language use, instantiating core claims of the

rich emergent theory of VWM which argues language comprehension and production processes support word memory, order memory, and their interaction (Schwering & MacDonald, 2020). By testing the model's performance on serial recall of memory lists and sentences, I can examine whether the specific computational architecture is equipped to account for recall of words vs nonwords; sentences vs lists; and sentences supported by the lexico-syntactic constraints of part-of-speech, verb biases, and animacy. The model described below, the Lichtheim-memory model, can perform all these tasks via its LTM.

**Modeling logic: General Approach and Ties to Rich Emergent Theory**

The Lichtheim-memory model expands upon a prior neurobiologically-inspired model of word repetition, comprehension, and production known as Lichtheim-2 (Ueno et al., 2011). Lichtheim-2 exhibits several valuable properties: it performs multiple tasks including comprehension, production, and repetition of words over a delay (i.e. recall); it learns to do these tasks through experience with language; and the learned representations integrate item (i.e. phone) and order (i.e. phonotactic) representations. Lichtheim-2 was specifically aimed at capturing properties of single word repetition, comprehension, and production. A small change to the model may allow it to extend to processing word sequences.

Lichtheim-2 is comprised of two main pathways, corresponding to dorsal and ventral streams that are thought to govern word comprehension and production in humans. By propagating activity through these pathways, the model divides the labor of word comprehension and production into two streams of phonological processing (dorsal) and semantic processing (ventral; see discussion in Ueno et al., 2011; Ueno et al., 2014). In humans, these same streams have been associated with higher order sentence processing, with the dorsal stream engaging in

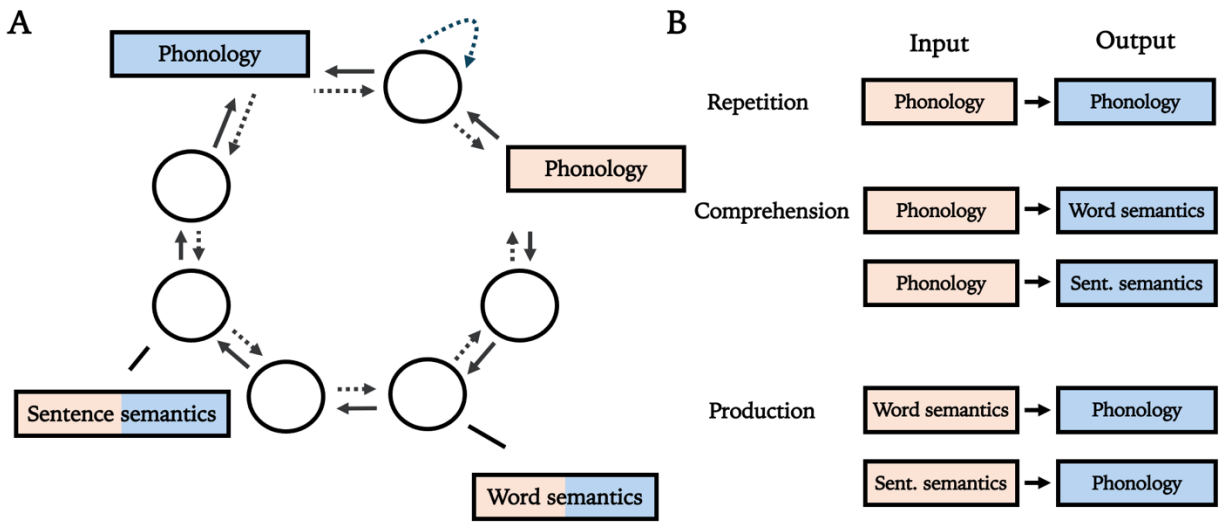
time-variant structure building and the ventral stream engaging in time-invariant event representation (Bornkessel-Schlesewsky & Schlesewsky, 2013).

The Lichtheim-memory model builds on Lichtheim-2 by adding a drive for the model to learn lexico-syntactic constraints. To do this, sentence semantic representations are integrated in the ventral stream. Furthermore, the Lichtheim-memory model is tasked with repeating, comprehending, and producing both words and sentences. For a visualization of the Lichtheim-memory model and the pathways governing different tasks, refer to *Figure 3*.

The proposed model captures gross properties of word comprehension and production. While the specific architecture was chosen with intention, several alternatives are viable and biologically plausible, like separable pathways for word and sentence semantics (e.g. Bornkessel-Schlesewsky & Schlesewsky, 2013). The current architecture was chosen to strike a balance between accuracy and tractability while capturing key interactions between word and sentence representations propagating among forward and backward connections.

**Figure 3**

*Lichtheim-memory model and tasks*




*Note.* A. The Lichtheim-memory model with input/output layers denoted by boxes, hidden layers denoted by circles, and weights connecting layers denoted by arrows. Forward connections denoted by solid arrows and recurrent connections through time denoted by dotted arrows. Inputs denoted by beige boxes and outputs denoted by blue boxes. B. The Lichtheim-memory model was trained on 3 different tasks: repetition, comprehension, and production. This schematic outlines the mapping from input to output sequences that the model must learn for each task. In comprehension and production tasks, the model received two sets of inputs and outputs: one set comprising word semantics and one set comprising sentence semantics.

## Method

The following sections define the Lichtheim-memory model, including the tasks on which the model was trained, the artificial language on which the model was trained, and the model architecture.

### *Tasks*

The Lichteim-memory model was trained to repeat, comprehend, and produce, akin to single word repetition, comprehension, and production tasks employed in Lichtheim-2 (Ueno et al., 2011). Unlike Lichtheim-2, the Lichtheim-memory model repeated, comprehended, and produced both single words and sentences.

**Repetition.** To repeat words and sentences, the model was presented a sequence of phonological inputs, after which the model was required to produce the same sequence as output. During presentation of the phonological input, the model was required to be silent, forcing the model to encode and maintain the entire input before repetition. During word repetition, the model received as input the phonological sequence of a single word and then repeated that phonological sequence. During sentence repetition, the model receivef as input the phonological sequence of a series of words and then repeated the entire phonological sequence of words. Each phoneme was be input to the model for 1 time step and then output for 1 timestep. It is important to note that, for this and all subsequent tasks, the model encountered no explicit marker of word boundaries. Sentences were comprised of concatenated word sequences, meaning the model needed to learn word boundaries through experience. In *Figure 1*, the repetition task required the model to map input from the *Phonological input* layer to output in the *Phonological output* layer.

**Comprehension**. To comprehend words and sentences, the model was provided a sequence of phonological inputs and required to produce a corresponding semantic representation. During word comprehension, the model was presented with the phonological sequence of one word. The model was required to immediately produce the word semantic representation of the presented word and maintain that semantic representation while the entire word was input to the model. During sentence comprehension, the model was presented with a

sequence of words. The model was then required to immediately produce the sentence semantic representation of the corresponding sentence and maintain that sentence semantic representation over the course of the sentence. During word comprehension, the model was not required to produce a sentence semantic representation, and vice versa. In *Figure 1*, the comprehension task required the model to map input from the *Phonological input* layer to output in the *Word semantics* layer or output in the *Sentence semantics* layer.

**Production.** During word production and sentence production tasks, the model was provided an input of word or sentence semantics, respectively. The model was then required to output a phonological sequence of the word or words expressing the semantics. In word production, the model was required to map input from the *Word semantics* layer to the *Phonological output* layer. In sentence production, the model was required to map input from the *Sentence semantics* layer to the *Phonological output* layer. During single word production, the model did not receive a sentence semantic input. Similarly, during sentence production, the model did not receive word semantic input.

### *Artificial language*

To force the model to learn lexico-syntactic constraints, the Lichtheim-memory model was trained on a carefully controlled artificial language. The artificial language was designed to reflect patterns in natural language, where lexico-syntactic constraints are known to impact both language use and VWM (Schwering et al., under review). The following section provides a high-level overview of the lexicon and grammar of the artificial language. In addition, the way the words and sentences of the artificial language are translated into phonological and semantic input and output patterns is described. Full details about the artificial language may be found in *Appendix A*.

**Artificial language: Words.** The lexicon of the model was composed of ditransitive verbs, transitive verb, intransitive verbs, animate nouns, and inanimate nouns. Phonology for words in the lexicon was constructed randomly using an onset consonant-vowel-offset consonant pattern (C-V-C). Semantic representations for words were defined by hand to capture a variety of features and differentiate words within the same roles in artificial sentences. A full list of phonological units comprising a word may be found in *Table A1*. A comparison of the artificial language semantic space and semantics of corresponding words in natural language may be found in *Figure A1*.

***Word phonology.*** Phonological representations of words were one-hot through time. The term "one-hot" means each phone corresponded to a single binary unit in a phonological feature vector. Phones were input through time such that the phonological representation of a word unfolds over multiple timesteps. When present in a word, a phone was turned on for 1 time step, meaning that feature received an input of 1 and all other features received an input of 0. The phonological feature vector comprised 12 phones, comprising 4 onset consonants, 4 vowels, and 4 offset consonants. Each word in the language was composed of 3 phones in a C-V-C pattern: an onset consonant, a vowel, and an offset consonant. Each word in the artificial language was randomly assigned a C-V-C pattern, and these patterns were held constant across all trained models. To see the full set of word phonology patterns, see *Table A1*.

***Word semantics.*** Semantic representations of words were copied from the localist semantic representations defined in the artificial language. When present for a word, a semantic feature was assigned an activation value of 1. Any feature not present was assigned a value of 0. Word semantic representations were time-invariant with respect to the word; the semantic representation of a word remained static during comprehension and production of that word. This

means that, during word comprehension, the model needed to map from a time-varying phonological input to a time-invariant semantic representation of the target word. During word production, the model needed to map from a time-invariant semantic representation to a time-varying phonological representation of the target word. Word semantic representations were of size 45, corresponding to the 45 localist semantic features defined in the artificial language.

**Artificial language: Grammar.** As in natural language, words of the lexicon were combined into sentences. Three types of sentences composed the artificial language: intransitive sentences (e.g. *boy blinked*), transitive sentences (e.g. *woman ate pizza*), and ditransitive sentences (e.g. *girl gave man gift*). A schematic of the rules governing the creation of sentences is visualized in *Figure A2*.

***Sentence phonology.*** Sentence phonology was generated by concatenating the phonological representations of words comprising that sentence. No explicit cue was provided to the model to indicate the boundary between words in a sentence. For example, during comprehension, the phonological input of the transitive sentence *man took letter* was input over 9 timesteps for the 3 phones in *man*, the 3 phones in *took*, and the 3 phones in *letter*.

***Sentence semantics.*** Representing sentence semantics poses a challenge for any model. While (non-contextual) word semantics may be clearly defined using localist representations, sentence semantics has no such analogue in the artificial language. Common solutions to representing sentence semantics in natural language include using contextualized word embeddings from language models trained to predict next or missing words from context (e.g. Devlin et al., 2018). However, such solutions are not easily tailored to the artificial language described above given the small size of the language. Further, contextualized word embeddings typically develop a different semantic representation at every time step during processing, which

may be inconsistent with language processing; some psycholinguistic language models ascribe the ventral language stream the objective of capturing time invariant semantic representations (Bornkessel-Schlesewsky & Schlesewsky, 2013).

To solve this issue, sentence semantic representations were made to be time invariant through training of latent sentence semantic representations in a separate model. Sentence representations were derived through implementation of a Sentence Gestalt model (St. John & McClelland, 1990) trained to respond to queries about sentences of the artificial language. As its name implies, the Sentence Gestalt model learns to represent a gestalt representation of the sentence, capturing elements of gist or event semantics. The Sentence Gestalt model has been shown to be a particularly good facsimile of language comprehension in humans, predicting N400 surprisal after being trained on both artificial languages (Rabovsky & McClelland, 2020) and natural language (Lopopolo & Rabvosky, 2021). Critically, the Sentence Gestalt captures semantic information about all elements in a sentence, given the latent sentence representation is employed to answer queries about all elements in the sentence. Implementational details of the Sentence Gestalt model developed for the Lichtheim-memory model may be found in *Appendix B*.

Sentence semantics were extracted from an independently trained Sentence Gestalt model. In this instantiation, the Sentence Gestalt model learned a latent representation of 10 units. Therefore, the extracted sentence semantic representations employed in the Lichtheim-memory model were comprised of a vector of 10 real-valued numbers (ranging in value from 0 to 1 due to sigmoidal activation in the Sentence Gestalt model). The sentence semantic representations were extracted after the final word in the sentence was input to the Sentence

Gestalt model to allow the sentence semantic representation to incorporate information across the entire sentence.

In sum, sentence semantic representations comprised a time invariant, real-valued vector capturing a distributed latent semantic representation of the event described in the sentence.

### *Model architecture*

The model was provided three input layers: input phonology, input word semantics, and input sentence semantics. Additionally, the model output its responses through three output layers: output phonology, output word semantics, and output sentence semantics. Activation of input layers and target activation of the output layers were pre-specified, defined by the task of the model and the representations of phonology, word semantics, and sentence semantics described above. The phonological input and output layers were of size 12, corresponding to the 12 phones in the artificial language. The word semantic input and output layers were of size 45, corresponding to the 45 word semantic features of the artificial language. The sentence input and output layers were of size 10, corresponding to the size of the gestalt layer in the Sentence Gestalt model trained on the artificial language.

Between input and output layers were hidden units, connected in sequence by learned weights. Weights fully connected adjacent layers in the network both forward and backward in time (see arrows in *Figure 3*). During processing, learned weights transformed input into output, instantiating the processing of language comprehension and production. Backward weights through time pass layer activity back to earlier layers in the processing stream at the next time step. While forward activity passes completely through the network in 1 time step, backward activity through time was propagated 1 layer per time step. This allows the model to integrate representations both across layers and through time. This means that, for any given timestep, a

layer needed to integrate both the current input and the latent representations the model has activated at previous timesteps. At the onset of a task, backwards connections were specified to have an input of 0.

All units, including output units, employed the sigmoid activation function. Phonological and semantic outputs ranged in value from 0 to 1.

### *Model training*

The Lichtheim-memory model was trained using the pytorch library, an open-source package designed to implement and train neural networks using the Python programming language. Training was performed on Google Colab. A total of 10 models were trained, each using an independent random seed, governing initialization of weights in the network and sampling of the artificial language.

For each model, the sentences of the artificial language were divided into training and testing sets, with a .75 train-test split. Sentences were sampled without replacement into the training set according to their probability. One important constraint governed sampling into the training set. To ensure a breadth of sentences were sampled into the training set, sentences were binned according to their structure, and sampling occurred with respect to these bins. Therefore, .75 of intransitive sentences, .75 of transitive sentences, and .75 of ditransitive sentences were sampled into the training set. During training, the model was exposed only to sentences in the training set.

Probabilities were assigned to individual sentences with respect to the probability of the structure of that sentence in rough proportion to natural English. Transitive sentences, being the most common in natural language, comprised .50 of the sentences in a training set. Ditransitive sentences, being less common in natural language, comprised .30 of the sentences in a training

set. Intransitive sentences, being the least common in natural language, comprised .20 of the sentences in a training set.

Training of the model followed the general procedure outlined in Ueno et al. (2011), with some modifications to train on both words and sentences. Training was divided into epochs. In each epoch, the model was exposed to a total of 300 sentences, with each sentence being presented 1 time in the context of repetition, 2 times in the context of comprehension, and 3 times in the context of production. The model was updated following exposure to each sentence using standard backpropagation. The order of sentences within each epoch was randomized.

In addition to training on full sentences, the model was tasked to complete single word repetition, single word comprehension, and single word production. Following each of the first 30 epochs, model exposure to words decreased stepwise. The model repeated, comprehended, and produced each word 3 times following the first 10 epochs. After epochs 11 through 20, the model repeated, comprehended, and produced each word 2 times. Then, after epochs 21 through 30, the model repeated, comprehended, and produced each word 1 time. Task frequency matched sentence training: when exposed to a word, the model repeated that word once, comprehended that word twice, and produced that word 3 times. Following the first 30 epochs of training, the model was no longer exposed to individual words, instead being trained solely on whole sentences.

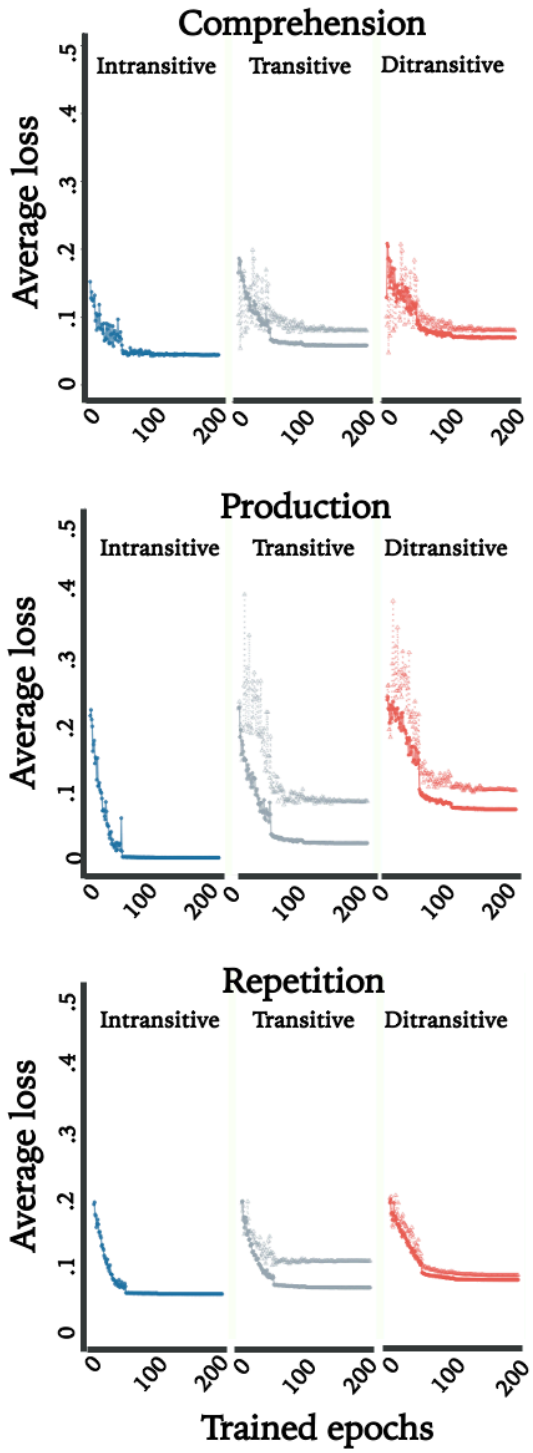
A total of 10 models were trained, each with a separate random seed governing weight initialization and sampling of sentences into training and testing sets. This was done to ensure evaluation of model performance was not dependent on specific seeds or training samples. For each model, weights were initialized randomly between -1 and 1, with biases initialized at 0.

Initial learning rate was set to 0.5 and decreased stepwise every 50 epochs with a decay gamma of .01. Weight decay was set to $1e^5$.

The total number of training epochs was determined empirically by examining training and testing loss on repetition, comprehension, and production of a separate set of pilot models. A priori, the total number of training epochs is challenging to specify. The rich language emergent theory would predict that memory capacity improves with language experience, in line with language modeling research which argues capacity for language comprehension and production improves with experience (e.g. MacDonald & Christiansen, 2002; Fitz et al., 2011). However, the relationship between one epoch and some amount of experience with language is indirect, often underspecified, and highly dependent upon task. For example, Gupta and Tisdale (2009) associated 1 training epoch with 1 year of experience with words, yet this was only a rough association. In the case of the Lichtheim-memory model, the highly constrained nature of the artificial language makes comparison to natural experience repeating, comprehending, and producing language challenging. Rather than pre-specifying a set amount of exposure, quantity of model training was set empirically given performance of pilot models. Loss on training and testing sets of the pilot models is visualized in *Figure 4*. A total of 200 epochs was chosen to capture the point at which training loss appeared to no longer decrease while minimizing overfitting on training data. All analyses of models were subsequently conducted on the models at the end of training.

**Figure 4**

*Training loss on comprehension, production, and repetition*



*Note.* Solid lines indicate loss on training sets. Lighter dotted lines indicate loss on testing sets. No lighter, dotted line is plotted for intransitive sentences due to sampling all intransitive sentences into training set for pilot models.

### *Testing model behavior*

In this project, I set out to test whether a rich, language emergent model of VWM could account for the ways in which linguistic LTM for words, sentences, and interactions among word and order representations influence VWM. To that end, my primary concern is the model's performance on repetition, an analogue to immediate serial recall. Model performance was tested through 5 tasks. In each task, the trained models were required to repeat novel lists, sentences, or sentence-like lists. These 5 tasks assess how the model's experience with language affects its word memory (Word regularity), order memory (Sentence-likeness), and its ability to integrate word and order representations through lexico-syntactic constraints. These 5 tasks are summarized in *Table 1*.

**Table 1**

*Tests of the Lichtheim-memory Model*

| Manipulation | Comparison | Example lists | Comparable behavioral study |
|---|---|---|---|
| Word regularity | Noun list<br>vs<br>Nonword list | sugar-man-pizza-girl<br>vs<br>frip-tog-wilp-stec | Hulme et al. (1995) |
| Sentence-likeness | Sentence<br>vs<br>Scrambled | man-gave-boy-sugar<br>vs<br>sugar-man-boy-gave | Allen et al. (2018) |
| Lexico-syntactic:<br>Part-of-speech | Ditransitive<br>vs<br>Noun | man-**gave**-boy-sugar<br>vs<br>man-**pizza**-boy-sugar | Schwering et al. (under rev.)<br>Experiment 1 |
| Lexico-syntactic:<br>Verb bias | Ditransitive<br>vs<br>Intransitive | man-**gave**-boy-sugar<br>vs<br>man-**slept**-boy-sugar | Schwering et al. (under rev.)<br>Experiment 2A |
| Lexico-syntactic:<br>Subject animacy | Animate<br>vs<br>Inanimate | **man**-gave-boy-sugar<br>vs<br>**pizza**-gave-boy-sugar | Schwering et al. (under rev.)<br>Experiment 2B |

*Note.* Each row representations a different test of the model. In the first 3 tests, the manipulation either comprises shuffling manipulating the word-likeness of all words in the list (Word regularity) or the order of the list (Sentence-likeness). In the last 3 tests, the manipulation

comprises 1 word, which is bolded in the column *Example lists*. Note, nonwords in the word regularity comparison represent rough approximations to the random C-V-C patterns used to test model repetition.

**Word superiority effect.** The first test compared repetition of 500 randomly generated lists of nouns with 500 randomly generated lists of nonwords. Nonwords were generated by combining all possible combinations of onset consonant, vowel, and offset consonant phones. From this set, all real words were removed, as well as words that violated phonotactic constraints in the artificial language (i.e. words that contained offset consonants or vowels in the onset consonant position; words that contained onset consonants or offset consonants in the vowel position, etc.). If the model generates a LTM sensitive to word phonotactic regularity and phonotactic regularity impacts word memory, then repetition of real words should be more accurate than repetition of nonwords.

**Sentence superiority effect.** The second test compared repetition of novel, well-formed ditransitive sentences from the test set with repetition of the same set of words presented in a random, different order. Scrambled sentences were generated by taking all well-formed ditransitive sentences and randomizing their order, removing from the set all valid sentences and removing all sentences that contained legal bigrams. This second restriction was put in place to limit minute effects of grammatical regularities (e.g. Perham et al., 2009). If the model generates a LTM sensitive to gross grammatical regularities and these gross grammatical regularities impact order memory, then repetition of the well-formed sentence should be more accurate than repetition of the same words in a scrambled order.

**Lexico-syntactic constraint: Part-of-speech.** The third test compared repetition of novel, well-formed ditransitive sentences with repetition of the same sentences, swapping out ditransitive verbs for nouns. If the model generate a LTM sensitive to the lexico-syntactic constraint of part-of-speech regularities and these same part-of-speech regularities impact repetition, then repetition of the sentence with the ditransitive verb should be more accurate than repetition of the same sentence with a noun in place of the ditransitive verb.

**Lexico-syntactic constraint: Verb biases.** The fourth test compared repetition of novel, well-formed ditransitive sentences with repetition of the same sentences, swapping out ditransitive verbs for intransitive verbs. If the model generates a LTM sensitive to the lexico-syntactic constraint of verb biases and these verb biases impact repetition, then repetition of the sentence with the ditransitive verb should be more accurate than repetition of the same sentence with an intransitive verb.

**Lexico-syntactic constraint: Subject animacy.** The fifth test compared repetition of novel, well-formed ditransitive sentences with repetition of the same sentences, swapping out the animate subject noun with an inanimate noun not otherwise present in the sentence. If the model generates a LTM sensitive to the lexico-syntactic constraint of subject animacy and these subject animacy impacts repetition, then repetition of the sentence with the animate subject noun should be more accurate than repetition of the same sentence with an inanimate subject noun.

## Results

Analyses of all models was conducted using generalized linear mixed effects regression predicting whether the most active output phone was the target output phone (1 = correct output, 0 = incorrect output) for each timestep. A summary of model performance on repetition tasks using this strict scoring procedure is provided in *Table 2*. Note, this scoring procedure differs

slightly from typical serial recall scoring procedures in the memory literature, which is typically conducted at the word level. To my knowledge, this scoring procedure is not standardized, as it often requires experimenter judgment as what constitutes a correct response given variations in accent or dialect, typos, errors, or other disfluencies.

In all cases, model performance was assessed using mixed effects logistic regression, predicting correct/incorrect responses from condition. Models and utterances were treated as random effects, with by-model and by-utterance random intercepts, as well as a by-model random slope for condition. The constrained sentence size of the artificial language limits comparison to positional effects and would be of limited theoretical value. Memory lists typically comprise 6 or more words; the artificial language has a maximum sentence size of 4 words. As a result, the primary model comparison concerns a gross, main effect of condition. The only comparison of interest of position was conducted in the analysis of the repetition of random noun lists and repetition of nonword lists, described below. Nevertheless, for all models, position and its interaction with condition were included in every model as a covariate.

**Table 2**

*Performance of the models*

| Stimulus type | Word 1 accuracy | Word 2 accuracy | Word 3 accuracy | Word 4 accuracy |
|---|---|---|---|---|
| Legal nonword list | 0.27 | 0.32 | 0.32 | 0.29 |
| Noun list | 0.42 | 0.44 | 0.48 | 0.57 |
| Test ditransitive sentences | 0.92 | 0.90 | 0.89 | 0.75 |
| Scrambled ditransitive sentences | 0.28 | 0.51 | 0.58 | 0.6 |
| Verb bias manipulation (inanimate noun) | 0.71 | 0.44 | 0.76 | 0.64 |
| Verb bias manipulation (transitive verb) | 0.62 | 0.47 | 0.77 | 0.70 |
| Verb bias manipulation (intransitive verb) | 0.68 | 0.56 | 0.76 | 0.67 |
| Subject animacy manipulation (inanimate noun) | 0.34 | 0.80 | 0.83 | 0.72 |
| Postverbal DO manipulation (inanimate noun) | 0.89 | 0.72 | 0.50 | 0.65 |

*Note.* Average repetition accuracy for all phones across conditions and models. Accuracy is given using strict serial scoring, rounded to 2 decimal places.

### *Word superiority effect*

In this comparison, repetition of legal nonword lists and noun lists were compared. There was a main effect of condition, such that lists of nouns were repeated better than lists of non-

words, $b = 0.84$, $X^2(1) = 105.15$, $p < .001$. Further, there was a main effect of list position on repetition, with repetition improving as list position increased, $b = 0.34$, $X^2(1) = 467.30$, $p < .001$. Finally, there was an interaction among condition and list position, indicating a greater improvement in repetition as position increased for lists of nouns compared to lists of non-words, $b = 0.50$, $X^2(1) = 253.14$, $p < .001$.

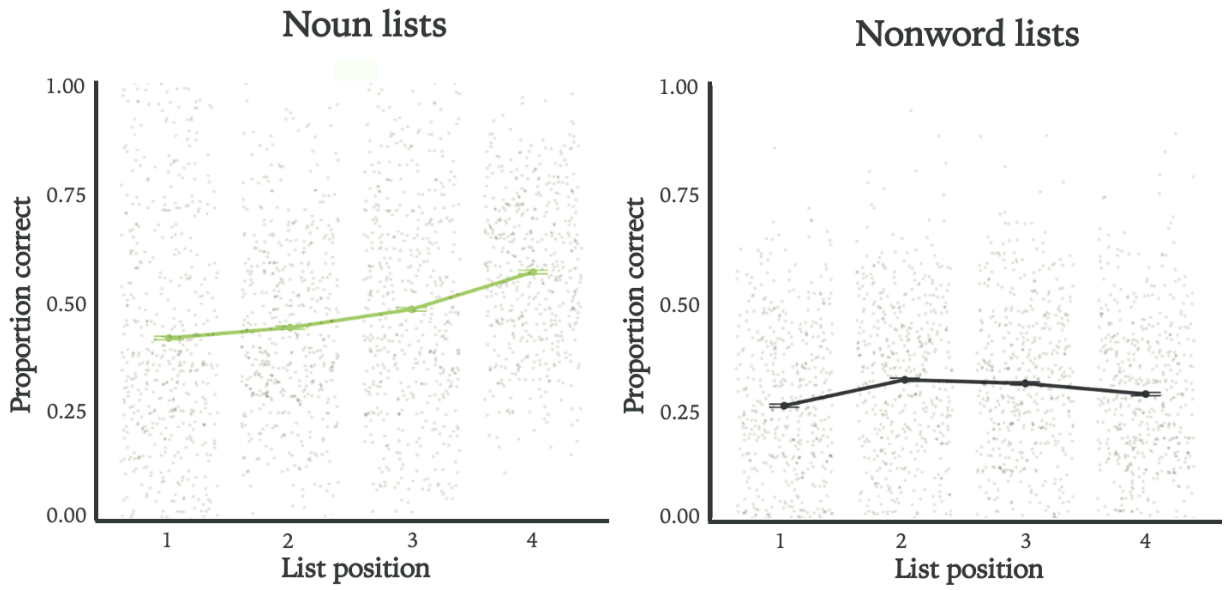
To further interrogate how repetition changed across position, an additional model was conducted examining the quadratic trend of position. The shape of serial position curves in serial recall includes primacy and recency effects, or improved recall for initial and final words in memory lists compared to middle list positions (Madigan, 1980). This finding is so foundational to VWM theorizing as to be a core target for models of VWM (Oberauer et al., 2018) and thus warrant a special analysis in noun lists typically presented in serial recall tasks. In linear regression, the presence of primacy or recency effects would be given by a significant effect of the quadratic trend of list position on recall, so an additional analysis was conducted examining this effect. This analysis included all previous fixed effects as well as the quadratic trend of position and interaction between the quadratic trend of list position and condition.

In this additional analysis, the Lichtheim-memory models were better at recalling lists of nouns over lists of non-words, as indicated by a main effect of condition, $b = 0.52$, $X^2(1) = 41.41$, $p < .001$. The models' performance increased as list position increased, regardless of condition, as indicated by a main effect of position, $b = 0.34$, $X^2(1) = 463.72$, $p < .001$. Additionally, repetition across position followed a quadratic trend, as indicated by a significant main effect of the quadratic trend of position, $b = -0.21$, $X^2(1) = 10.15$, $p < .01$. The effect of the quadratic trend differed across conditions, as indicated by an interaction between condition and the quadratic trend of position, $b = 1.83$, $X^2(1) = 194.02$, $p < .001$. Visual inspection of the trend revealed a

slight recency effect in the noun list, as indicated by improved repetition in final list positions, that was absent in the non-word list. There was no primacy effect. Finally, there was also an interaction between condition and linear list position, $b = 0.50$, $X^2(1) = 246.95$, $p < .001$, indicating the change in repetition across list position differed between conditions. A visualization of repetition performance can be seen in *Figure 5*.

**Figure 5**

*Repetition of non-word and noun lists*




*Note.* Small dots indicate jittered repetition accuracy for each word in an utterance collapsed across all trained models. All bars represent standard error.

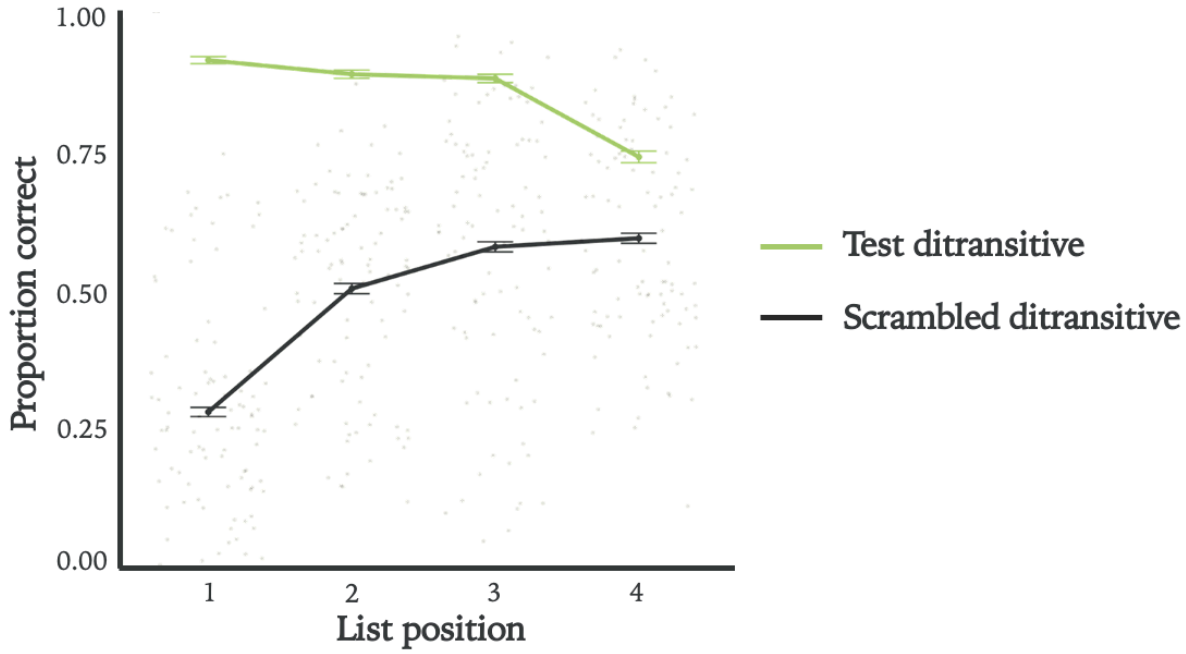
### *Sentence superiority effect*

In this comparison, repetition of novel, test ditransitive sentences and scrambled ditransitive sentences were compared. Analyses revealed the models were better at recalling novel, well-formed ditransitive sentences compared to scrambled ditransitive sentences, as

indicated by a main effect of condition, $b = 1.06$, $X^2(1) = 693.48$, $p < .001$. This effect was consistent across all positions, though it was particularly pronounced in the initial list position. There was no significant main effect of list position on repetition, $X^2(1) = 1.24$, $p = .27$, though there was an interaction between condition and list position, $b = 2.54$, $X^2(1) = 529.87$, $p < .001$. Visualization of scrambled sentence repetition is visualized in *Figure 6*.

**Figure 6**

*Repetition of scrambled sentences*




*Note*. Black dots and solid lines indicate performance on ditransitive sentences with the violated lexico-syntactic constraint. Green lines indicate performance on test ditransitive sentences. Small dots indicate jittered repetition accuracy for each word in an utterance collapsed across all trained models. All bars represent standard error.

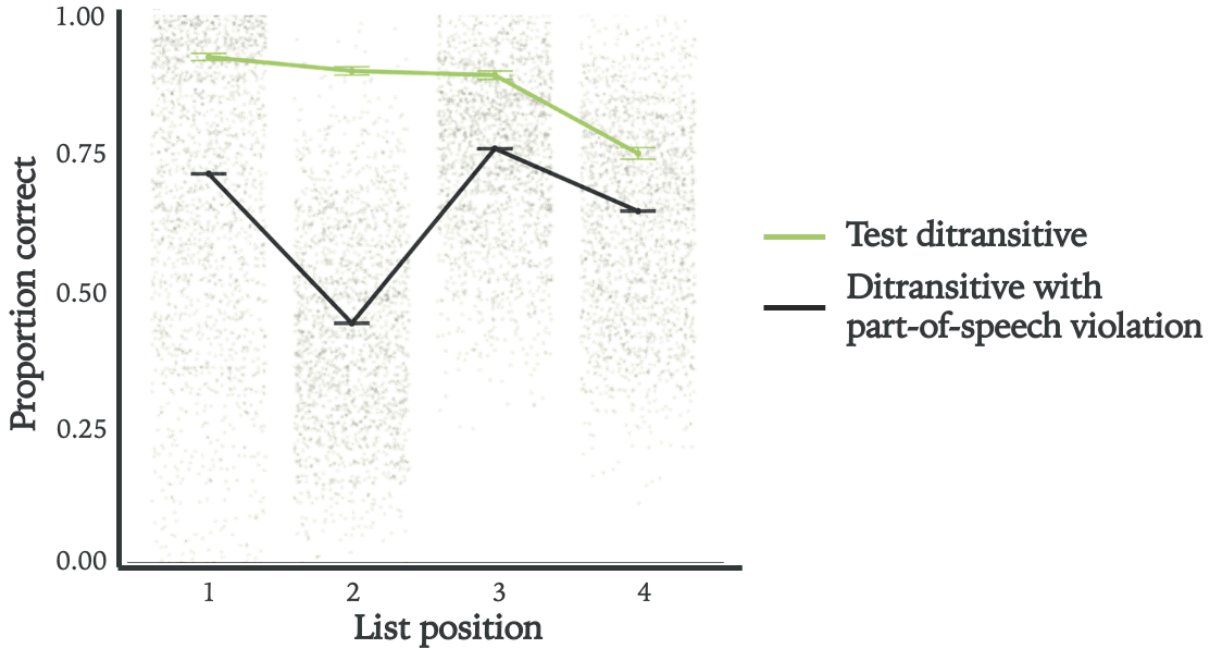
***Lexico-syntactic constraint: Part-of-speech***

In this comparison, repetition of novel, test ditransitive sentences and ditransitive sentences with the ditransitive verb swapped with an inanimate noun were compared. Repetition was better for well-formed ditransitive sentences compared to ditransitive sentences with a part-of-speech violation in which the ditransitive verb was replaced with an inanimate noun, $b = 1.54$, $X^2(1) = 320.40$, $p < .001$. Furthermore, there was an effect of list position, such that repetition decreased as list position increased, $b = -0.50$, $X^2(1) = 116.05$, $p < .001$. Finally, there was an

interaction between condition and list position, such that the decrement in repetition was greater for well-formed sentences than sentences with a part-of-speech violation, $b = -1.67$, $X^2(1) = 327.89$, $p < .001$. Repetition of sentences with a part-of-speech violation is visualized in *Figure 7*.

**Figure 7**

*Repetition of sentences with and without violated part-of-speech regularity constraint*




*Note.* The black solid line and dots indicate performance on ditransitive sentences with the violated lexico-syntactic constraint. The green line indicates performance on test ditransitive sentences. Small dots indicate jittered repetition accuracy for each word in an utterance collapsed across all trained models. All bars represent standard error.

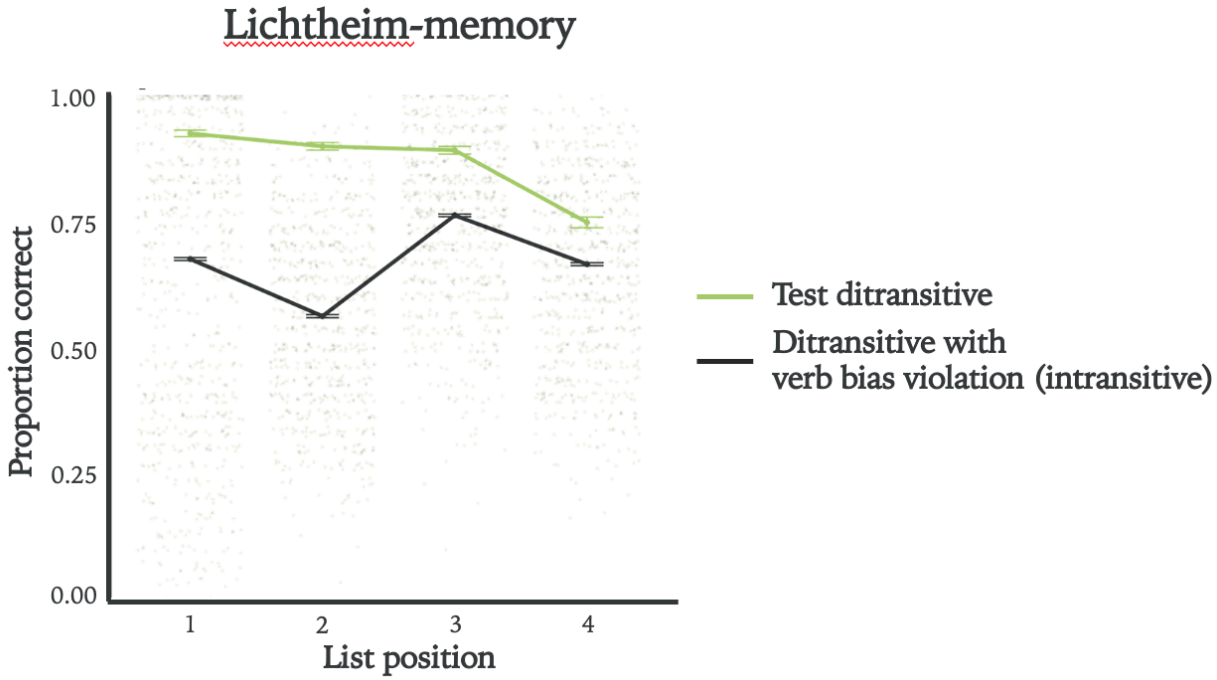
### *Lexico-syntactic constraint: Verb biases*

Two separate models were fit to analyze how replacing a ditransitive verb in a novel, well-formed test ditransitive sentence affected repetition performance. Replacing a ditransitive verb with a transitive or intransitive verb similarly impaired repetition of sentences. Both models had a significant main effect of condition. Repetition was impaired when the ditransitive verb was replaced with a transitive verb, $b = 1.54$, $X^2(1) = 308.55$, $p < .001$, and when the ditransitive verb was replaced with an intransitive verb, $b = 1.33$, $X^2(1) = 269.44$, $p < .001$. Furthermore, in

both models, performance decreased across list position, as indicated by a main effect of list position when a transitive verb replaces the ditransitive verb, $b = -0.25$, $X^2(1) = 24.84$, $p < .001$, and when an intransitive verb replaces the ditransitive verb, $b = -0.50$, $X^2(1) = 99.36$, $p < .001$. Finally, in both models, there was an interaction between condition and list position, both for the transitive verb manipulation, $b = -2.19$, $X^2(1) = 488.29$, $p < .001$, and the intransitive verb manipulation, $b = -1.69$, $X^2(1) = 289.07$, $p < .001$. Repetition of sentences with violated verb biases is visualized in *Figures 8* and *9*. Repetition with a swapped intransitive verb is visualized in *Figure 8*, and repetition with a swapped transitive verb is visualized in *Figure 9*.

**Figure 8**

*Repetition of sentences with and without violated verb bias constraint (intransitive)*




*Note.* The black solid line and dots indicate performance on ditransitive sentences with the violated lexico-syntactic constraint. The green line indicates performance on test ditransitive sentences. Small dots indicate jittered repetition accuracy for each word in an utterance collapsed across all trained models. All bars represent standard error.

**Figure 9**

*Repetition of sentences with and without violated verb bias constraint (transitive)*


*Note.* The black solid line and dots indicate performance on ditransitive sentences with the violated lexico-syntactic constraint. The green line indicates performance on test ditransitive sentences. Small dots indicate jittered repetition accuracy for each word in an utterance collapsed across all trained models. All bars represent standard error.

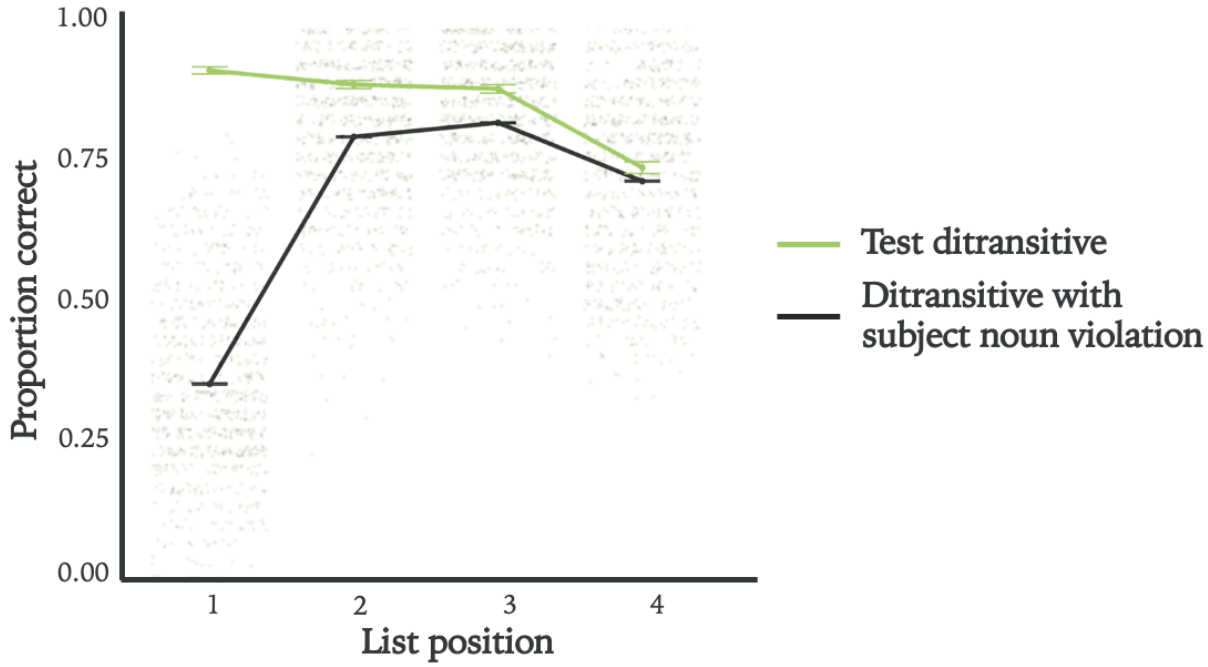
***Lexico-syntactic constraint: Subject animacy***

In this comparison, repetition of novel, test ditransitive sentences and ditransitive sentences with the subject noun replaced with an inanimate noun were compared. The Lichtheim-memory models were less accurate at repeating ditransitive sentences with an animate subject noun than an inanimate subject noun, $b = 1.20$, $X^2(1) = 165.11$, $p < .001$. Further, as list position increased, repetition slightly improved, $b = 0.10$, $X^2(1) = 4.31$, $p < .05$. Finally, there

was an interaction among condition and list position such that repetition decreased across list position for well-formed sentences but not sentences with an inanimate subject, $b = -2.86$, $X^2(1) = 964.80$, $p < .001$. Repetition of sentences with violated subject noun animacy is visualized in *Figure 10*.

**Figure 10**

*Repetition of sentences with and without violated subject animacy constraint*




*Note.* The black solid line and dots indicate performance on ditransitive sentences with the violated lexico-syntactic constraint. The green line indicates performance on test ditransitive sentences. Small dots indicate jittered repetition accuracy for each word in an utterance collapsed across all trained models. All bars represent standard error.

## Discussion

In this project, I developed and tested the Lichtheim-memory model, an instantiation of the rich emergent theory of VWM (Schwering & MacDonald, 2020). This model was tested on a series of serial recall tasks and compared with human behavioral data to determine whether language comprehension and production processes could plausibly subserve VWM. In accordance with human subjects data, the model captured 5 ways in which linguistic LTM supports VWM: improved memory for words over nonwords (Hulme et al., 1995), for sentences over scrambled

sentences (Allen et al., 2018), for sentence-like lists with consistent part-of-speech patterns over inconsistent patterns, for sentence-like lists with consistent verb biases over inconsistent verb biases, and for sentence-like lists with animate subjects over inanimate subjects (Schwering et al., under review). Critically, each performance pattern was driven entirely by the model's linguistic LTM, trained to complete the tasks of language comprehension, production, and repetition. Together, these results provide a plausible computational underpinning to a mechanism by which VWM can emerge from processing language.

**Comparisons to prior memory theories and models**

Emergent theories of VWM starkly contrast dominant, buffer theories of VWM (Baddeley, 2000; Baddeley & Hitch, 1974; Norris, 2017). According to buffer theories, VWM is a distinct cognitive module, temporarily storing and manipulating memoranda to be used by other cognitive systems. In contrast, language emergent theories argue language comprehension and production processes support encoding, maintenance, and retrieval of linguistic memoranda. A bevy of behavioral and neuroscientific research supports this perspective (Acheson & MacDonald, 2009; Schwering & MacDonald, 2020; MacDonald, 2016). Nevertheless, emergent theory has been stymied by a lack of computational models, leading proponents of buffer theories to argue emergent theory is untenable. Critics of emergent theories have argued emergent computational models are either non-existent or wholly incapable of accounting for hallmark phenomenon in the VWM literature (Norris, 2017). Generating an emergent computational model is challenging; in addition to developing a model of VWM, emergent theorists need consider both constraints on language use and VWM. The Lichtheim-memory model provides one instantiation of emergent VWM. In this model, language processing and linguistic LTM is characterized as a transformation of phonological and semantic representations

in the service of comprehension, production, and repetition in a simple recurrent neural network. Temporary memory is characterized as the activation of the language system in the service of comprehension, production, and repetition.

Models of VWM have rarely adopted perspectives from language research. For many buffer theories, VWM is an explicit cognitive module. For example, the Start-end model (Henson, 1998) generates a memory trace through an explicit, endogenously generated ordering signal. Of course, models need not incorporate an explicit ordering signal, and some models blur the line between LTM and VWM without explicitly adopting emergent VWM theory. The Context Retrieval and Updating model (Logan, 2018; Logan & Cox, 2021) and the Temporal Context Model (TCM, Howard and Kahana, 2002) both build temporary memory from fading traces of earlier presented items and their contexts. In these models, particularly the TCM, LTM can be accounted for by incorporating interference or support from traces over long delays. Nevertheless, ties to specific language comprehension and production mechanisms remain sparse. In contrast, the Lichtheim-memory model is an explicit instantiation of the mechanisms of language comprehension, production, and repetition.

Unlike buffer models, however, the Lichtheim-memory model was trained solely to comprehend, produce, and repeat sentences in an artificial language and not to explicitly repeat real words or consistent sentences better than non-words or inconsistent sentences. That is, the Lichtheim-memory model was not explicitly fit to the target behavioral benchmarks. Instead, the Lichtheim-memory model's capacity to repeat sentences is a function of its experience. Prior research into language modeling suggests this is a natural product of neural networks. Capacity of language models improves with training, leading experienced models to better map on to adult language comprehension and production than less experienced models (e.g. MacDonald &

Christiansen, 2002; Fitz et al., 2011). The Lichtheim-memory model demonstrates a similar pattern in the domain of VWM. Alternative models of VWM employ a very different set of techniques. To account for improved performance of memoranda stored in linguistic LTM, buffer theories make reference to redintegration (e.g. Jones & Farrell, 2018). Over time, more modules have been added to the memory models to account for the ways in which different types of linguistic LTM can affect processing, such as the way in which part-of-speech patterns enhance recall in Jones and Farrell (2018). In comparison to buffer models, the Lichtheim-memory model provides a parsimonious account of both language processing and VWM.

The Lichtheim-memory model shares several similarities with other emergent models of VWM but expands their processing to integrate word and sentence representations. Prior emergent models have characterized how phonological, lexical, and word semantic representations support memory for words. For example, the lexical network of Martin et al. (1996) employed spreading activation between phonological, lexical, and semantic word representations to capture word memory in healthy and aphasic populations. While this model could capture single word performance, the model did not incorporate relations among words into its processing, leaving its application to the VWM literature regarding list memory ambiguous. In another vein, Botvinick and Plaut (2006) created a simple recurrent neural network and trained it to perform serial recall. While this model was not explicitly linked to language use, it could accurately capture several VWM phenomenon, including transposition effects (Henson et al., 1996) and primacy and recency effects (Madigan, 1980). The Lichtheim-memory models combines the best of both works, applying a neurobiologically inspired language network that incorporates both word and word order representations in the service of VWM tasks.

**Tested behavioral phenomena and future benchmarks**

The Lichtheim-memory model provides a clear way to account for the sentence superiority effect. The sentence superiority effect is the finding that sentence-like lists tend to be recalled better than scrambled sentences (e.g. Allen et al., 2018). This finding has been established at multiple grain sizes, with researchers finding improved memory for attested phrases over random strings (Arnon & Snider, 2010; Jacobs et al., 2016) for canonical adjective-noun sequences over the reversed (Perham et al., 2009l Schweppe et al., 2022), and noun-noun compounds consistent with patterns in natural language over the reversed (Schwering & MacDonald, under review). The sentence superiority effect is important for the memory literature, not only to establish that linguistic LTM influences VWM performance, but also because it suggests that LTM retains a specific character. Namely, these patterns suggest that word and word order representations interact to influence VWM; sentence structure is a function of inter-word relationships, taking on meaning both as a function of the individual words themselves but also their context (see Schwering & MacDonald, 2020 for discussion). This is critical for theories of VWM, which typically characterize word or "item" representations as wholly distinct from word order representations (Majerus, 2013). In contrast to this memory perspective, the Lichtheim-memory model adopts the rich emergent theory of VWM which does not consider word and word order representations to be wholly distinct (Schwering & MacDonald, 2020). In simple recurrent neural networks like the Lichtheim-memory model, representations from prior timesteps bear on the present, naturally integrating word and word order representations. In Lichtheim-2, phonological and semantic representations were largely but not wholly separable along ventral and dorsal streams (Ueno et al., 2011; Ueno et al., 2014). Similarly, it is expected that the Lichtheim-memory model should incorporate word and word

order representations throughout processing. The fact the Lichtheim-memory model could comprehend, produce, and repeat sentences is initial evidence that this information influenced behavior. Future research, discussed below, will consider this explicitly.

There are several ways in which the Lichtheim-memory model's performance could better match human performance. First, with respect to the current tests, the model fails to capture the subtle ways in which lexico-syntactic constraints shape VWM when verb biases are manipulated. In human behavioral data, participants' recall performance tracked the relative goodness of sentence-like lists. For example, when recalling ditransitive sentences with an intransitive verb replacing the ditransitive verb, participants' performance appeared to drop when the sentence context was no longer supported by the verb (see *Figure 2*, Experiment 2A). Consider the following case: when recalling the sentence-like list HAPPY-DOG-SLEPT-ANGRY-BUTCHER-PAPER, participants recalled the partial sentence HAPPY-DOG-SLEPT at similar rates to a well-formed ditransitive sentence. However, they recalled the unwarranted phrase ANGRY-BUTCHER-PAPER relatively poorly compared to the well-formed ditransitive sentence. The Lichtheim-memory model did not track this sensitivity to partial well-formedness of the sentence, instead recalling the entire sentence-like list poorly (see *Figure 8* and *Figure 9*). Further interrogation of the model's sentence semantics representations should reveal whether repetition of violated lexico-syntactic constraints show a graded degradation, or whether the overall decrement in performance is matched by catastrophic degradation of sentence semantic representations.

Additionally, there are several memory benchmarks that need be considered to establish the Lichtheim-memory model as a general model of VWM. Oberauer et al (2018) identified 67 benchmarks for computational models of VWM. While no model of VWM can currently capture

all benchmarks, several core benchmarks of high priority need be considered. Many core benchmarks would likely require alteration of the model's training tasks or architecture: set-size effects on accuracy (Unsworth & Engle, 2006) and reaction time (Mayberry et al., 2002), impairments in recall in presence of concurrent cognitive load (Vallar & Baddeley, 1982), improved recall for grouped memoranda (Hitch et al., 1996), preservation of short-term memory in amnesia (Baddeley & Warrington, 1970; though see Dell et al., 1997), and so on. Other phenomenon are viable target benchmarks using the existing architecture due to their reliance on linguistic LTM or establishment in previous simple recurrent neural networks: effects of chunking and number of remembered chunks (Miller & Selfridge, 1950; Botvinick & Bylsma, 2005), effects of relatedness of memoranda on confusions (Henson et al., 1996; Saint-Aubin & Poirier, 1999), locality constraints on transpositions (Henson et al., 1996; Botvinick & Plaut, 2006), phonological similarity effects (Conrad & Hull, 1964), and so on. Nevertheless, in its current instantiation, the Lichtheim-memory model provides a useful account of the sentence superiority effect and word superiority effects in an emergent framework, an important extension of prior memory models.

Primacy and recency effects deserve additional interrogation, given, at first glance, the model's ability to capture recency effects in repeating random lists of nouns. Primacy and recency effects, the pattern of improved recall for list initial and list final memoranda, are considered core phenomenon in VWM tasks (Madigan, 1980). Typically, primacy effects are functionally captured through enhanced repetition of list initial items (Page & Norris, 1998; Tan & Ward, 2008), and recency effects are captured through minimal decay for recently encountered memoranda (Page & Norris, 1998), though researchers have proposed a number of alternative mechanisms (e.g. Oberauer & Lewandowsky, 2008). While the Lichtheim-memory

model demonstrates a recency effect in random memory lists composed of nouns, the source of these effects is not currently clear. This may be a function of minimal interference between memoranda, given list final nouns are encountered immediately before output. However, these effects may also be driven by the structure of the model's training data; all transitive and ditransitive sentences encountered during training end in nouns, meaning the model's experience at later list positions was comprised solely of nouns. This would make nouns relatively frequent in list final positions compared to earlier list positions, resulting in a primacy effect in noun lists. Nevertheless, this hypothesis is undermined by the fact that repetition of transitive and ditransitive sentences does not also exhibit a recency effect. Indeed, visual inspection of repetition of sentences reveals what appears to be a small primacy effect. Further research could be done using models trained on varying artificial languages to determine whether the effect is reliable or an artifact of the language.

**Applications to the language literature**

Many language comprehension and production theories characterize VWM as a constraint on language use. How does the Lichtheim-memory model fit into this framework? Rather than considering VWM as a separate cognitive constraint on language processing, the Lichtheim-memory model characterizes VWM as an emergent capacity from language comprehension and production. This perspective aligns with constraint-satisfaction theories of language processing (Seidenberg, 1997; MacDonald & Seidenberg, 2006), which emphasize the role multiple interacting sources of linguistic LTM play in shaping language use. Indeed, language users tend to produce more familiar syntactic structures and structures supported by lexical constraints, even when the less familiar syntactic structures induce a lower VWM demand (Montag & MacDonald, 2015). Similarly, in the Lichtheim-memory model, the relative

frequency of different structures influences processing, making well-formed sentence structures easier to repeat than random memory lists or sentences with violated structures. Rather than emphasizing the importance of a separate VWM capacity on language processing, the Lichtheim-memory model emphasizes the importance of language experience. In this way, the Lichtheim-memory model continues a long line of using computational models, and neural networks, in particular, to characterize language processing (e.g. Joanisse & Seidenberg, 2003; Joanisse & McClelland, 2015).

Computational modeling and the Lichtheim-memory model could provide one avenue to compare memory-limited (e.g. Lewis et al., 2006) and constraint-satisfaction theories (e.g. MacDonald & Seidenberg, 2006). In its current state, the Lichtheim-memory model provides an account of the ways in which language experience can predict language use. One could imagine adding a separate VWM capacity to the network to see whether the fit to human performance improves with a separate VWM capacity. In parallel, Hahn et al. (2022) applied a similar approach to large language models, demonstrating human sentence comprehension and production patterns could be predicted better using both measures of memory load and sentence surprisal than memory load or surprisal alone. Similarly, the Lichtheim-memory model could provide one avenue to test theories of language use by comparing comprehension and production of sentences in the model's current form and when a separate VWM capacity is added to the model. The addition of a separate VWM capacity could take several forms. Lewis et al. (2006) have characterized both the importance of decay and interference in governing language use. The current version of the Lichtheim-memory model incorporates interference in the form of multiplexed signals in the model's activated state, and decay could easily be incorporated as a

function of time (e.g. Page & Norris, 2009). Regardless, the Lichtheim-memory model could serve as a means of providing computational specificity to VWM in language models.

**Future directions**

There are several ways in which the model's performance could be improved. Of foremost concern is improving the model's training experience. Developing a more naturalistic language on which the model could be trained as well as a more naturalistic training procedure should improve the extent to which the model would map on to human performance. The current artificial language is heavily restricted, consisting of sentences with a maximum of 4 words, with many classes of words missing, like articles or adjectives. Including alternative sentence structures would be critical to testing the role that language experience and VWM plays in language processing. Additionally, expansion of the model's training set would allow testing of additional memory phenomenon. The sentence superiority effect has been extended to adjective noun pairs (Perham et al., 2009) and noun compounds (Schwering & MacDonald, under review), which could prove a fruitful avenue for testing the model's sensitivity to sentence-likeness at multiple grain sizes. Additionally, it is worth noting that the model's ability to extend to novel examples is a function of the breadth of its training experience. A relatively restricted set of training examples means the model's generalizability to novel examples is restricted. Memory lists, typically lists of nouns, do not reflect the sentences in the artificial language. However, in natural language, ill-formed speech and lists of nouns appear with some frequency, requiring the language comprehension and production system adapt to many examples that may better reflect memory lists. Extending the model's training environment to include ill-formed utterances and lists of nouns may allow the model to better capture language phenomenon as well as VWM.

The psycholinguistic literature provides a rich set of mechanisms that can be used to further refine the Lichtheim-memory model. One particularly interesting area of consideration is that of neural oscillations. Neural oscillations have been linked to structure building mechanisms in language processing, with oscillations entrained to detection of phones, combination of phones into words, and in phrasal processing (Meyer, 2018). These neural oscillations have been explicitly linked to cognitive mechanisms that allow the model to process novel sentences through compositional structure building (Martin & Doumas, 2017). Though the extension of these mechanisms to a simple recurrent neural network remains unclear, and is at odds with neural networks (Martin & Doumas, 2017), it is worth noting that oscillations are often invoked in the VWM literature as a form of order maintenance. For example, the Primacy gradient (Page & Norris, 1998; 2009) and Start-end models (Henson, 1998) encode memory as activation strengths from an oscillating order signal. While ties to language mechanisms remains sparse, there is some research tying psycholinguistic oscillations to the memory literature. The bottom-up multi-scale population oscillator model (BUMP) was inspired by acoustic signal processing (Hartley et al., 2016). BUMP employs battery of oscillators to encode a continuous memory signal into item- and list-level representations. The psycholinguistic literature typically characterizes oscillations as an endogenously generated grouping signal; incorporating oscillations as either a learned or architectural constraint into the Lichtheim-memory model could serve as one way to improve the model's structure-building abilities and link the model closer with prior work in the memory literature.

Several additional analyses have yet to be conducted to understand the mechanisms supporting processing in the Lichtheim-memory model. The latent sentence semantic representations are a particular area of interest. These sentence semantic representations are

hypothesized to support sentence processing in repetition. If this is the case, then lesioning of these representations should result in degradation of sentence repetition performance. Beyond establishing the causal role of the sentence semantic representations in repetition, lesioning the model could be a particularly fruitful way of modeling impairments of language function and recovery (Ueno et al., 2011). Classic aphasiology has argued that language processing and VWM may be governed by distinct modules (e.g. Baddeley & Warrington, 1970), though VWM impairments are often comorbid with aphasia (Caplan & Waters, 1995). The Lichtheim-memory model provides an opportunity to computationally capture sentence comprehension, production, and repetition impairment as a function of damage to the different components of the model. Additionally, given the model's emphasis on learning LTM representations, modeling recovery training regimes via sentence repetition could be a potential target application (e.g. Eom & Sung, 2016).

## References


Acheson, D. J., & MacDonald, M. C. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological Bulletin, 135*(1), 50–68. https://doi.org/10.1037/a0014411

Acheson, D. J., MacDonald, M. C., & Postle, B. R. (2011). The effect of concurrent semantic categorization on delayed serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 44–59. https://doi.org/10.1037/a0021205

Acheson, D. J., Postle, B. R., & MacDonald, M. C. (2010). The interaction of concreteness and phonological similarity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 17–36. https://doi.org/10.1037/a0017679

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools, 49*(3), 340–355. https://doi.org/10.1044/2018_LSHSS-17-0114

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science*, *17*(9), 814-823.

Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2018). Exploring the sentence advantage in working memory: Insights from serial recall and recognition. Quarterly Journal of Experimental Psychology, 71(12), 2571–2585. https://doi.org/10.1177/1747021817746929

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, *62*(1), 67-82.

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences*, *4*(11), 417-423.

Baddeley, A. D. (2017). Modularity, working memory and language acquisition. Second Language Research, 33(3), 299–311. https://doi.org/10.1177/0267658317709852

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47-89). Academic press.

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language, 61*(3), 438–456. https://doi.org/10.1016/j.jml.2009.05.004

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological review*, *105*(1), 158.

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning & Verbal Behavior, 9*(2), 176–189.

https://doi.org/10.1016/S0022-5371(70)80048-2

Berg, T. (2014). On the relationship between type and token frequency. *Journal of Quantitative Linguistics*, *21*(3), 199-222.

Botvinick, M., & Bylsma, L. M. (2005). Regularization in Short-Term Memory for Serial Order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 351–358. https://doi.org/10.1037/0278-7393.31.2.351

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological review*, *113*(2), 201.

Caplan, D., & Waters, G. S. (1995). Aphasic disorders of syntactic comprehension and working memory capacity. *Cognitive Neuropsychology, 12*(6), 637–649. https://doi.org/10.1080/02643299508252011

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology, 55*(4), 429–432. https://doi.org/10.1111/j.2044-8295.1964.tb00928.x

Cowan, N. (1993). Activation, attention, and short-term memory. *Memory & cognition*, *21*(2), 162-167.

Cowan N. (2008). What are the differences between long-term, short-term, and working memory?. *Progress in brain research*, *169*, 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Crowder, R. G. (1993). Short-term memory: Where do we stand? *Memory & Cognition, 21*(2), 142–145. https://doi.org/10.3758/BF03202725

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eom, B., & Sung, J. E. (2016). The Effects of Sentence Repetition-Based Working Memory Treatment on Sentence Comprehension Abilities in Individuals With Aphasia. *American journal of speech-language pathology*, *25*(4S), S823–S838. https://doi.org/10.1044/2016_AJSLP-15-0151

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(32), 12203–12208. https://doi.org/10.1073/pnas.0602173103

Fitz, H., Chang, F., Christiansen, M. H., & Kidd, E. (2011). A connectionist account of the acquisition and processing of relative clauses. *The acquisition of relative clauses*, *8*, 39-60.

Freedman, M. L., Martin, R. C., & Biegler, K. (2004). Semantic relatedness effects in conjoined noun phrase production: Implications for the role of short-term memory. *Cognitive Neuropsychology*, *21*(2-4), 245-265.F

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*(2), 200–213. https://doi.org/10.1016/0749-596X(89)90044-2

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*(2), 228–244. https://doi.org/10.1016/0749-596X(90)90074-A

Gupta, P., & Tisdale, J. (2009). Word learning, phonological short-term memory, phonotactic probability and long-term memory: towards an integrated framework. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3755-3771.

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(43), e2122602119. https://doi.org/10.1073/pnas.2122602119

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. https://doi.org/10.1016/j.cogpsych.2016.05.001

Henson, R. N. A. (1998). Short-term memory for serial order: The Start–End Model. *Cognitive Psychology, 36*(2), 73–137. https://doi.org/10.1006/cogp.1998.0685

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. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 49A*(1), 80–115. https://doi.org/10.1080/027249896392810

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 49A*(1), 116–139. https://doi.org/10.1080/027249896392829

Hsiao, Y., & MacDonald, M. C. (2016). Production predicts comprehension: Animacy effects in Mandarin relative clause processing. *Journal of Memory and Language*, *89*, 87-109.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Hulme, C., Roodenrys, S., Brown, G., & Mercer, R. (1995). The role of long-term memory mechanisms in memory span. *British Journal of Psychology, 86*(4), 527–536. https://doi.org/10.1111/j.2044-8295.1995.tb02570.x

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1217–1232. https://doi.org/10.1037/0278-7393.23.5.1217

Jacobs, C. L., Dell, G. S., & Bannard, C. (2017). Phrase frequency effects in free recall: Evidence for redintegration. *Journal of memory and language*, *97*, 1-16.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, *87*, 38-58.

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *WIREs Cognitive Science, 6*(3), 235–247. https://doi.org/10.1002/wcs.1340

Joanisse, M. F., & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and language*, *86*(1), 40-56.

Johns, B. T. (2021). Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity. *Memory & Cognition*, 1-20.

Jones, T., & Farrell, S. (2018). Does syntax bias serial order reconstruction of verbal short-term memory? *Journal of Memory and Language, 100,* 98–122. https://doi.org/10.1016/j.jml.2018.02.001

Kowialiewski, B., Lemaire, B., Majerus, S., & Portrat, S. (2021). Can activated long-term memory maintain serial order information? *Psychonomic Bulletin & Review, 28*(4), 1301–1312. https://doi.org/10.3758/s13423-021-01902-3

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

Logan, G. D. (2018). Automatic control: How experts act without thinking. *Psychological Review, 125*(4), 453–485. https://doi.org/10.1037/rev0000100

Logan, G. D., & Cox, G. E. (2021). Serial memory: Putting chains and position codes in context. *Psychological Review, 128*(6), 1197–1205. https://doi.org/10.1037/rev0000327

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of memory and Language*, *31*(6), 713-733.

Lopopolo, A., & Rabovsky, M. (2021). Predicting the N400 ERP component using the Sentence Gestalt model trained on a large scale corpus. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43. Retrieved from https://escholarship.org/uc/item/49x8z7bm

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9*(2), 157–201. https://doi.org/10.1080/01690969408402115

MacDonald, M. C. (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. https://doi.org/10.1177/0963721415620776

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35–54. https://doi.org/10.1037/0033-295X.109.1.35

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive psychology*, *24*(1), 56-98.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703. https://doi.org/10.1037/0033-295X.101.4.676

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In *Handbook of psycholinguistics* (pp. 581-611). Academic Press.

Madigan, S. (1980). The serial position curve in immediate serial recall. *Bulletin of the Psychonomic Society, 15*(5), 335–338. https://doi.org/10.3758/BF03334550

Majerus S. (2013). Language repetition and short-term memory: an integrative framework. *Frontiers in human neuroscience*, *7*, 357. https://doi.org/10.3389/fnhum.2013.00357

Majerus, S. (2018). Working memory treatment in aphasia: A theoretical and quantitative review. *Journal of Neurolinguistics*, *48*, 157-175.

Martin, A. E., & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS biology*, *15*(3), e2000663. https://doi.org/10.1371/journal.pbio.2000663

Martin, R. C., & Freedman, M. L. (2001). Short-term retention of lexical-semantic representations: Implications for speech production. *Memory*, *9*(4-6), 261-280.

Martin, N., Saffran, E. M., & Dell, G. S. (1996). Recovery in deep dysphasia: Evidence for a relation between auditory–verbal STM capacity and lexical errors in repetition. *Brain and Language*, *52*(1), 83-113.

Martin, R. C., Shelton, J. R., & Yaffee, L. S. (1994). Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. *Journal of Memory and Language*, *33*(1), 83-111.

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*(3), 360–385. https://doi.org/10.1016/S0749-596X(02)00014-1

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*(3), 283–312. https://doi.org/10.1006/jmla.1997.2543

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience, 48*(7), 2609–2621. https://doi.org/10.1111/ejn.13748

Miller, G. A., & Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *The American Journal of Psychology, 63,* 176–185. https://doi.org/10.2307/1418920

Monaghan, P., & Woollams, A. M. (2017). Implementing the "Simple" model of reading deficits: A connectionist investigation of interactivity. In *Neurocomputational Models of Cognitive Development and Processing: Proceedings of the 14th Neural Computation and Psychology Workshop* (pp. 69-81).

Nikravesh, M., Aghajanzadeh, M., Maroufizadeh, S., Saffarian, A., & Jafari, Z. (2021). Working memory training in post-stroke aphasia: Near and far transfer effects. *Journal of communication disorders*, *89*, 106077. https://doi.org/10.1016/j.jcomdis.2020.106077

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin, 143*(9), 992-1009. http://dx.doi.org/10.1037/bul0000108

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review, 115*(3), 544–576. https://doi.org/10.1037/0033-295X.115.3.544

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin, 144*(9), 885–958. https://doi.org/10.1037/bul0000153

Page, M., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological review*, *105*(4), 761.

Page, M. P., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1536), 3737–3753.

https://doi.org/10.1098/rstb.2009.0173

Perham, N., Marsh, J. E., & Jones, D. M. (2009). Syntax and serial recall: How language supports short-term memory for order. *The Quarterly Journal of Experimental Psychology, 62*(7), 1285–1293. https://doi.org/10.1080/17470210802635599

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 50*(4), 408–412. https://doi.org/10.1037/1196-1961.50.4.408

Portrat, S., Barrouillet, P., & Camos, V. (2008). Time-related decay or interference-based forgetting in working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1561–1564. https://doi.org/10.1037/a0013356

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23-38.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633-654.

Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *375*(1791), 20190313. https://doi.org/10.1098/rstb.2019.0313

Roodenrys, S., Hulme, C., & Brown, G. (1993). The development of short-term memory span: Separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology, 56*(3), 431–442. https://doi.org/10.1006/jecp.1993.1043

Saint-Aubin, J., & Poirier, M. (1999). Semantic similarity and immediate serial recall: Is there a detrimental effect on order information? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 52A*(2), 367–394. https://doi.org/10.1080/027249899391115

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223–250). Lawrence Erlbaum Associates, Inc.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27*(5), 499–520. https://doi.org/10.1016/0749-596X(88)90022-8

Schwering, S. C., & MacDonald, M. C. (2020). Verbal Working Memory as Emergent from Language Comprehension and Production. *Frontiers in human neuroscience*, *14*, 68. https://doi.org/10.3389/fnhum.2020.00068

Schwering, S. C., Jacobs, Jacobs, C. L., Montemayor, J., MacDonald, M. C. (under review). Lexico-syntactic properties affect verbal working memory in sentence-like lists.

Schweppe, J., Schütte, F., Machleb, F., & Hellfritsch, M. (2022). Syntax, morphosyntax, and serial recall: How language supports short-term memory. *Memory & Cognition.* Advance online publication. https://doi.org/10.3758/s13421-021-01203-z

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275*(5306), 1599–1603. https://doi.org/10.1126/science.275.5306.1599

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science, 23*(4), 569–588. https://doi.org/10.1016/S0364-0213(99)00016-6

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Slevc, L. R., & Martin, R. C. (2016). Syntactic agreement attraction reflects working memory processes. *Journal of Cognitive Psychology*, *28*(7), 773-790.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Spivey-Knowlton, M., & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, *55*(3), 227–267. https://doi.org/10.1016/0010-0277(94)00647-4

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 47*(2), 276–309. https://doi.org/10.1037/h0078826

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial intelligence*, *46*(1-2), 217-257.

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density.

Storkel, H. L., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken American English. *Behavior research methods*, *42*(2), 497-506.

Szewczyk, J. M., & Schriefers, H. (2011). Is animacy special?: ERP correlates of semantic violations and animacy violations in sentence processing. *Brain research*, *1368*, 208-221.

Tan, Y., Martin, R. C., & Van Dyke, J. A. (2017). Semantic and Syntactic Interference in Sentence Comprehension: A Comparison of Working Memory Models. *Frontiers in psychology*, *8*, 198. https://doi.org/10.3389/fpsyg.2017.00198

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review, 15*(3), 535–542. https://doi.org/10.3758/PBR.15.3.535

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. *Architectures and mechanisms for language processing*, 90-118.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental psychology: Learning, memory, and Cognition*, *19*(3), 528.

Ueno, T., Saito, S., Rogers, T. T., & Ralph, M. A. L. (2011). Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(2), 385-396.

Ueno, T., Saito, S., Saito, A., Tanida, Y., Patterson, K., & Lambon Ralph, M. A. (2014). Not lost in translation: Generalization of the primary systems hypothesis to Japanese-specific language processes. *Journal of cognitive neuroscience*, *26*(2), 433-446.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language, 54*(1), 68–80. https://doi.org/10.1016/j.jml.2005.06.003

Vallar, G., & Baddeley, A. D. (1982). Short-term forgetting and the articulatory loop. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 34A*(1), 53–60. https://doi.org/10.1080/14640748208400857

Van Dyke, J. A., & Johns, C. L. (2012). Memory Interference as a Determinant of Language Comprehension. *Language and linguistics compass*, *6*(4), 193–211. https://doi.org/10.1002/lnc3.330

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481-487.

Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1256–1271. https://doi.org/10.1037/0278-7393.25.5.1256

Wright, H. H., & Shisler, R. J. (2005). Working memory in aphasia: theory, measures, and clinical implications. *American journal of speech-language pathology*, *14*(2), 107–118. https://doi.org/10.1044/1058-0360(2005/012)

# Appendix A

## Artificial language

The artificial language was designed to include the sentence structures and semantic features targeted in all tests of model performance. The following describes the lexicon of the artificial language and provides an overview of the grammatical rules governing construction of sentences in the language.

### *Lexicon and semantic features*

At minimum, the lexicon of the artificial language needed to comprise of (1) a set of verbs that differed in their transitivity, (2) a set of nouns that could serve as subjects in all sentences, and (3) a set of nouns that could serve as direct and indirect objects in transitive sentences. While addressing these three grammatical constraints, loose semantic constraints were also considered. Semantic features were chosen to provide a wide array of semantic features that would correlate with certain events and not others while keeping the language no larger than necessary to test effects of verb biases on processing. For example, a range of foods with different properties were chosen to occur with *fed* or *ate* sentences but not *mailed* or *drank* verbs. Ultimately, a total of 4 intransitive, 4 transitive, and 4 ditransitive verbs were selected, along with 4 possible subject nouns, and 11 possible direct objects. Word semantic features were designed by hand, with some of the words and features borrowed from Rabovsky & McClelland (2018).

Word phonetic and semantic features are listed in *Table A1*. *Figure A1* provides a visual summary of the cosine similarity of the semantic features of the words in the artificial language and in natural language. There was no explicit objective to generate a set of semantic features

that would re-create a real semantic space, though the comparison does demonstrate some overlap between the artificial space and the real.

***Grammar***

The grammar of the artificial language needed to contain, at minimum, two structures: (1) ditransitive sentences with two objects of the verb and (2) intransitive sentences. In the current iteration of the artificial language, a third structure, expressing simple transitive sentences with only one object of the verb, was also included, to add a degree to uncertainty to the model's predictions.

All sentences in the artificial grammar were constructed through the verb. Verbs assigned probabilities to both the sentence structures they afforded as well as the fillers that may fill the roles required by the chosen sentence structure. Note, in natural language, verbs, sentences structures, and fillers all mutually inform one another. These constraints do not inform this artificial language. The construction of an example sentence in the artificial language is illustrated in *Figure A2*.

**Table A1**

*Word representations in the artificial language*

| Word | Role | Phonological features | Semantic features |
|---|---|---|---|
| blinked | Verb, intransitive | C1-3, V1-3, C2-0 | action, body, blinked |
| ran | Verb, intransitive | C1-3, V1-0, C2-2 | action, body, ran |
| sat | Verb, intransitive | C1-1, V1-0, C2-2 | action, body, sat |
| slept | Verb, intransitive | C1-2, V1-1, C2-1 | action, body, slept |
| ate | Verb, transitive | C1-1, V1-1, C2-0 | action, meals, body, ate |
| drank | Verb, transitive | C1-1, V1-2, C2-1 | action, meals, body, drank |
| took | Verb, transitive | C1-1, V1-2, C2-2 | action, social, took |
| borrowed | Verb, transitive | C1-2, V1-0, C2-2 | action, social, borrowed |
| fed | Verb, ditransitive | C1-1, V1-3, C2-2 | action, meals, fed |
| gave | Verb, ditransitive | C1-3, V1-2, C2-2 | action, social, gave |
| lent | Verb, ditransitive | C1-2, V1-0, C2-1 | action, social, lent |
| mailed | Verb, ditransitive | C1-1, V1-0, C2-3 | action, social, mailed |

| | | | |
|---|---|---|---|
| man | Noun, S, IO | C1-3, V1-0, C2-3 | person, active, adult, male, man |
| woman | Noun, S, IO | C1-1, V1-0, C2-0 | person, active, adult, female, woman |
| boy | Noun, S, IO | C1-0, V1-3, C2-2 | person, active, male, boy |
| girl | Noun, S, IO | C1-3, V1-2, C2-1 | person, active, female, girl |
| eggs | Noun, DO | C1-2, V1-2, C2-1 | consumable, food, white, eggs |
| toast | Noun, DO | C1-0, V1-0, C2-1 | consumable, food, brown, toast |
| pizza | Noun, DO | C1-2, V1-3, C2-3 | consumable, food, pizza |
| coffee | Noun, DO | C1-1, V1-1, C2-2 | consumable, drinkable, brown, coffee |
| tea | Noun, DO | C1-0, V1-3, C2-0 | consumable, drinkable, tea |
| soup | Noun, DO | C1-2, V1-2, C2-2 | consumable, drinkable, food, soup |
| sugar | Noun, DO | C1-3, V1-0, C2-0 | consumable, food, white, sugar |

| | | | |
|---|---|---|---|
| book | Noun, DO | C1-2, V1-3, C2-1 | writing, book |
| letter | Noun, DO | C1-0, V1-3, C2-1 | writing, mail, letter |
| package | Noun, DO | C1-3, V1-3, C2-1 | mail, package |
| gift | Noun, DO | C1-2, V1-1, C2-0 | social, birthday, gift |

*Note.* The *role* column indicates role in which a word can occur in the artificial language. A tag of S indicates that the word occurs as a subject. A tag of IO or DO indicates that the word occurs as an indirect object or a direct object, respectively. The *phonological features* column indicates the phonemes of the word, in sequence. Phones were randomly assigned to words based on an artificial C-V-C pattern (C1 = onset consonant; V1 = vowel; C2 = offset consonant). For example, a pattern of C1-0, V1-1, C2-3 indicates the word contained the first onset consonant, the second vowel, and the fourth offset consonant. The *semantic features* column indicates what semantic features were assigned to the word. All words had at least 1 unique semantic feature, as indicated by the word being repeated in the *semantic features* column.

**Figure A2**

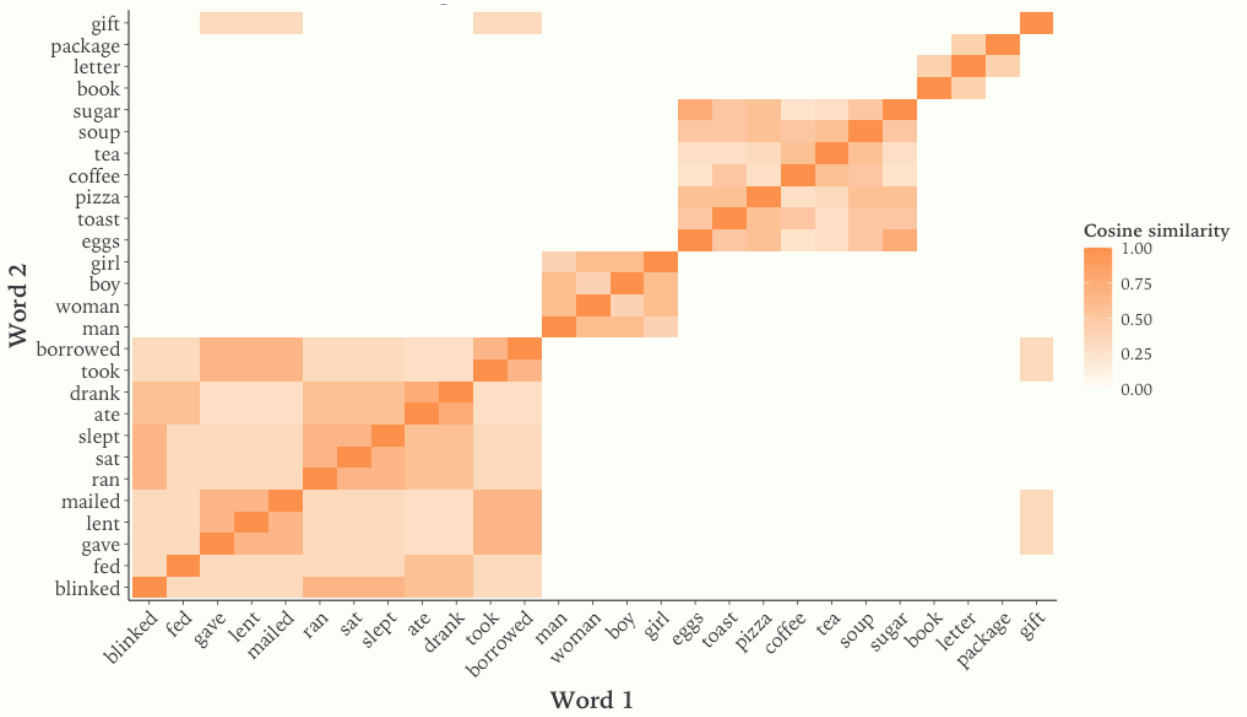
*Verb biases in the artificial language*

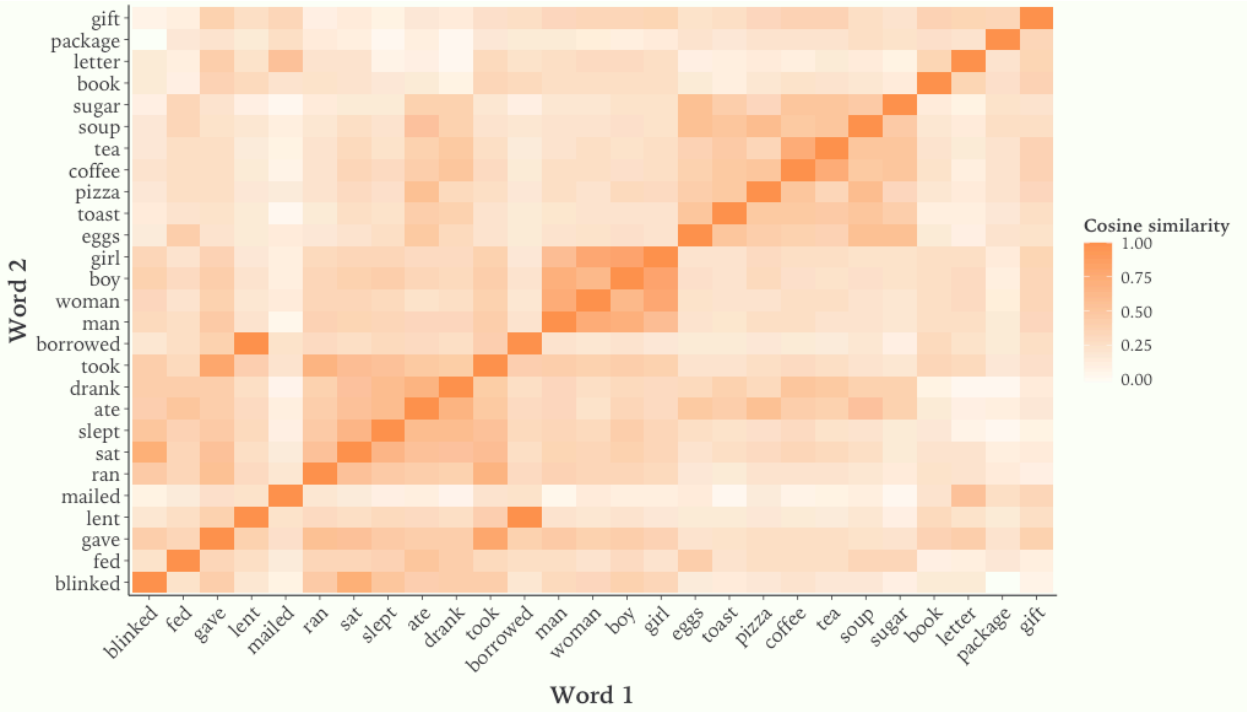
| Verb | Dominant construction | *p*(Intransitive) | *p*(Transitive) | *p*(Ditransitive) |
|---|---|---|---|---|
| blinked | Intransitive | 1 | 0 | 0 |
| ran | Intransitive | 1 | 0 | 0 |
| sat | Intransitive | 1 | 0 | 0 |
| slept | Intransitive | 1 | 0 | 0 |
| ate | Transitive | 0 | 1 | 0 |
| drank | Transitive | 0 | 1 | 0 |
| took | Transitive | 0 | 1 | 0 |
| borrowed | Transitive | 0 | 1 | 0 |
| fed | Ditransitive | 0 | 0.1 | 0.9 |
| gave | Ditransitive | 0 | 0.1 | 0.9 |
| lent | Ditransitive | 0 | 0.1 | 0.9 |
| mailed | Ditransitive | 0 | 0.1 | 0.9 |

*Note.* Verb biases. Probability that a verb occurs in a specific construction in the training set for the Sentence Gestalt model. Note, ditransitive verbs could occasionally appear in intransitive sentences.

**Figure A1**

*Cosine similarities between semantic representations of words*

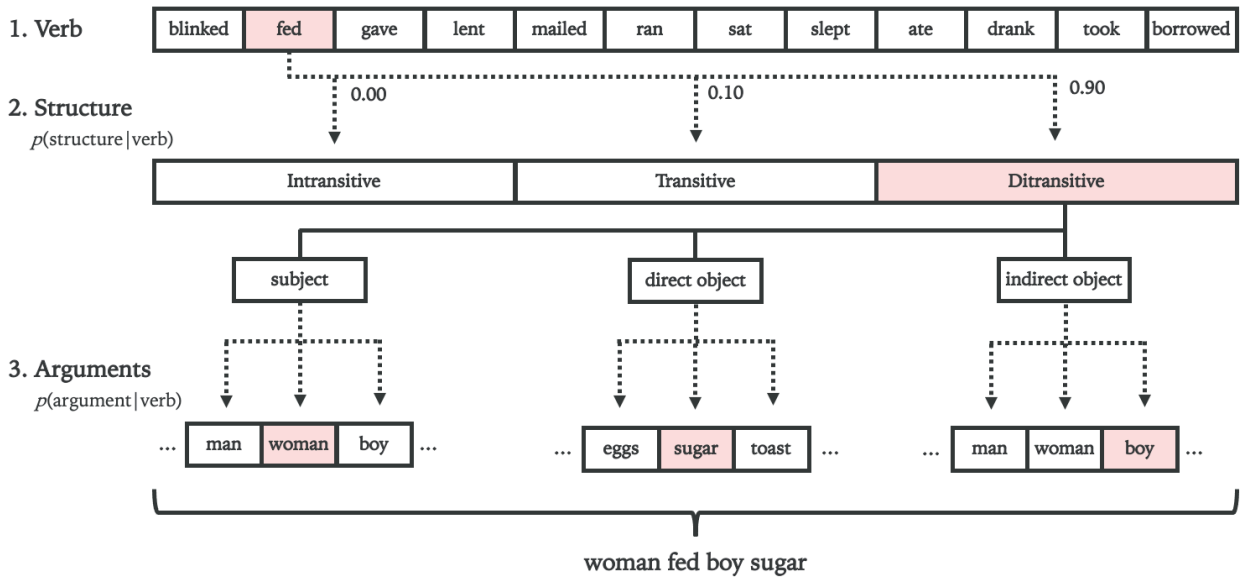


*Note.* The top figure displays the similarities between the hand-defined localist representations. The bottom figure displays the similarities between word embeddings taken from a pre-trained *en_core_web_md* language model in the spacy language processing python package.

**Figure A2**

*Construction of sentences in the artificial language*




*Note*. Construction of an example sentence in the artificial language. Example sentences are generated in the artificial language through the verb. After a verb is selected, a valid syntactic structure is selected. In this example, the verb *fed* licenses two potential syntactic structures: a transitive sentence with probability .10 and a ditransitive sentence with probability .90. Next, fillers are selected to fill the roles required by this syntactic structure. In this example, the selected ditransitive syntactic structure requires a subject, direct object, and indirect object. The probability of each filler (probabilities not pictured) is again governed by the verb. To produce the final sentence, the verb and the selected fillers are slotted into the syntactic structure.

# Appendix B

## Sentence Gestalt model

The Sentence Gestalt model, as its name implies, learns a *gestalt* representation of the event described in a sentence. The model learns to do this following exposure to a series of word inputs, in which the model needs to answer questions about the role/filler pairs of the sentence. The model must do this task even for words it has not yet encountered, meaning that the model must learn dependencies between words and their contexts. For example, if the model were exposed to the incomplete sentence *The woman gave*…, it could be queried on both known information, like the subject, as well as currently unknown information, like the upcoming direct object.

While the details of the currently instantiated Sentence Gestalt model are described below, reference to foundational work by St. John & McClelland (1990), contemporary ties to human neuroimaging data by Rabovsky & McClelland (2018), as well as training on large, naturalistic corpora by Popolo & Rabovksy (2021) will be informative for the curious reader.

### *Model architecture*

The Sentence Gestalt model is composed of a word input, a probe input, a role/filler output, and hidden layers connecting these inputs and outputs. The critical layer of the model is the sentence gestalt hidden layer, which contains recurrent connection through time to another hidden layer. This layer develops a representation of the event in the sentence as it unfolds in time. Critically, this sentence gestalt is forced to be generated by the model prior to integration with probes into the sentence. As a result, the model is forced to learn a representation of an event described in a sentence that can express information about all role/filler pairs in that sentence.

The current trained model was comprised of an input layer of size 31 units (27 words + 3 timing units + 1 padding unit), a hidden layer preceding the sentence gestalt of size 10 units, a second hidden layer labelled the sentence gestalt layer of size 10 units, a probe input layer of size 50 units (4 roles + 45 semantic features + 1 padding unit), a hidden layer integrating input from the probe and the sentence gestalt of size 10 units, and a final output layer of size 50 (same as the probe input). Layers were connected by weights (randomly initialized between -1 and 1), with biases (initialized at 0). All layers used a sigmoid activation function. An illustration of the model and the specific manner in which layers were connected may be seen in *Figure B1*.

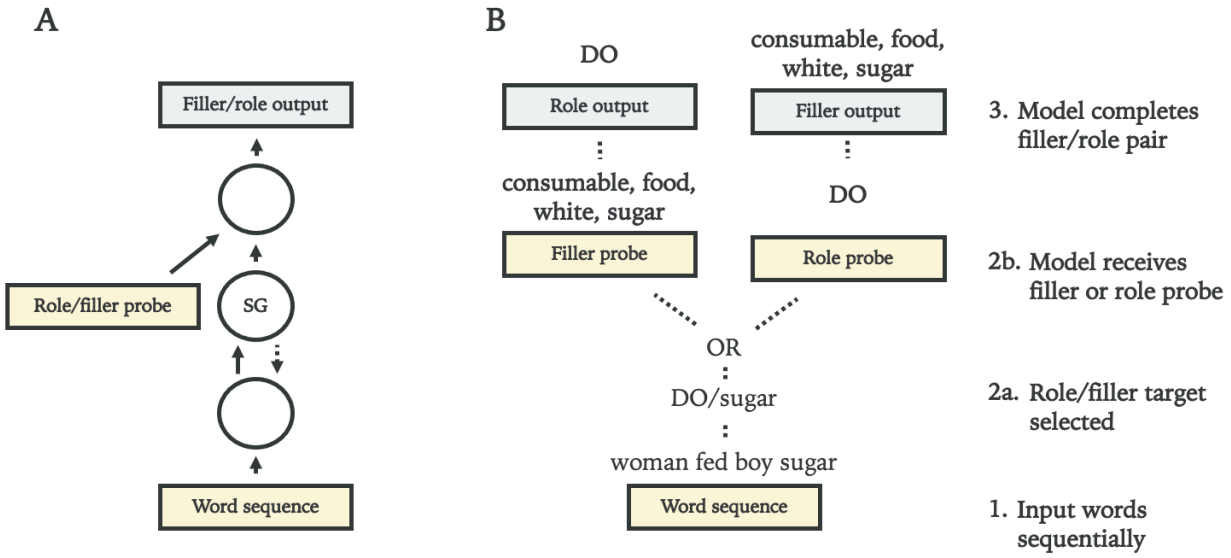
***Model task and training procedure***

In a training example, the model receives two inputs. The first is a series of one-hot encoded words concatenated with a timing signal indicating the time of a word relative to the verb in the sentence (see St. John & McClelland, 1990). The second is role/filler probe composed of either a one-hot representation of the target role, or a localist representation of the semantic features of the target filler. Together, these two sources of information inform the model's behavior as it attempts to complete the target role/filler pair indicated by the probe.

The Sentence Gestalt model was trained on all possible sentences of the artificial language described in *Appendix A*. For each sentence, the model was trained on all possible probes into role/filler pairs of that sentence. The model was trained on 10000 epochs of all sentences in the artificial language. Batch size was set to 64. Learning rate was initially set to .01 with a decay gamma of .01. Decay of the learning rate was step-wise every third of the way through training (i.e. epoch 0 $lr$ = .01; epoch 3333 $lr$ = .001; epoch 6666 $lr$ = .0001). Loss was calculated using binary cross entropy, and weights were updated using stochastic gradient descent, with momentum of .1.

**Figure B1**

*Architecture of the Sentence Gestalt Model with Schematic of Training Example*

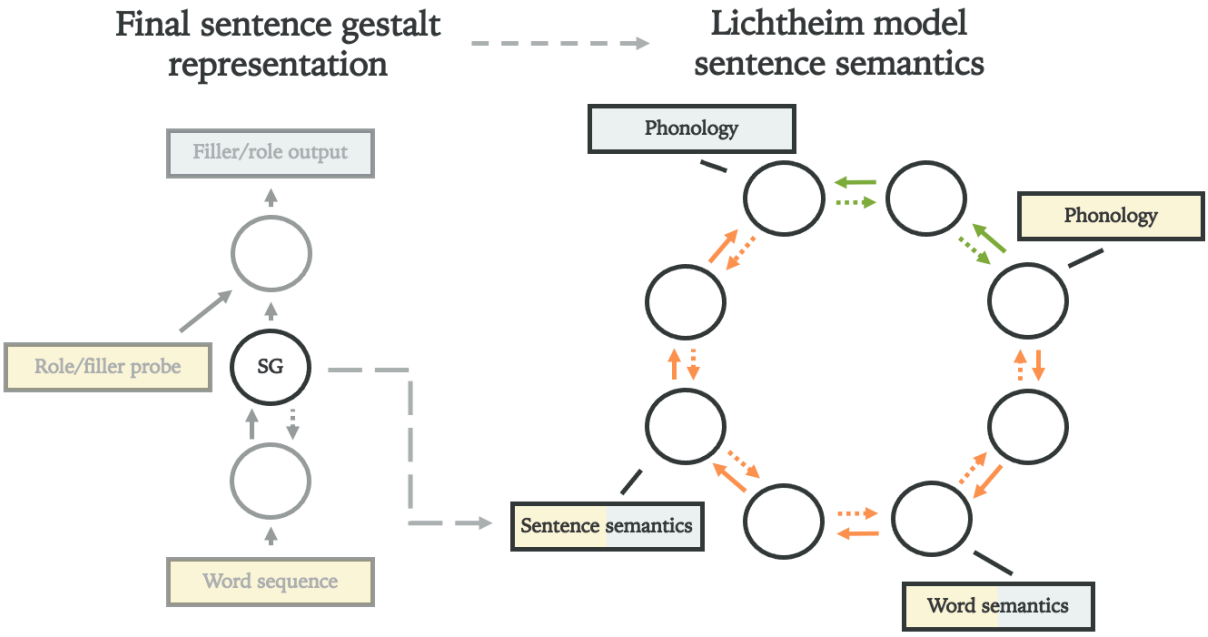



*Note.* Architecture of the Sentence Gestalt model with schematic of training example. In part *A*, forward projecting weight are denoted by solid lines while recurrent connections through time are denoted by the dotted line. Inputs are denoted by beige boxes and outputs denoted by blue boxes. Hidden layers are denoted by circles. Note, the hidden layer labeled SG comprised the *sentence gestalt* representation learned by the model. It is of particular importance the model generates this representation without reference to the role/filler probe. As a result, the *sentence gestalt* representation is the same for any probe, and the model must learn a *sentence gestalt* representation that can "answer" the query of any probe on which it is trained. In part B, two possible training examples are presented for the sentence *woman fed boy sugar*. These examples were chosen to illustrate the training procedure; the model is probed both on the roles in the sentence (e.g. the direct object sugar), and the fillers (e.g. features of the direct object sugar). In either case, the model must output the missing component of the role/filler pair. When presented with a role probe, the model must output the corresponding filler information of the word that

fulfills that role. When presented with a filler probe, the model must output the corresponding role that word fills. Note, the model tries to complete the role/filler pair at every time step with every presented word, even during the time steps before which the model has encountered the role/filler.

**Figure B2**

*Relationship between the Sentence Gestalt Model and the Lichtheim-memory Model*




*Note.* Relationship between the Sentence Gestalt model and the Lichtheim-memory model. Note the Lichtheim-memory model is simplified to conserve space; the sentence gestalt representation is fed into the sentence semantics layer of the full Lichtheim-memory model. Sentence gestalt representations from the layer labelled SG are used as target and input for comprehension and production tasks, respectively. The sentence gestalt representation employed in the Lichtheim-memory model is taken from the Sentence Gestalt model after the final word is presented, meaning that the Lichtheim-memory model's task of mapping from phonology to semantics entails comprehending semantic features of words that may not yet be encountered. While sentence gestalt representations develop over the course of exposure to a sentence, the Lichtheim-memory model will treat the input/output semantic representation of the word sequence as time-invariant.