RUNNING HEAD: Statistically-based chunking of non-adjacent dependencies

Revision of XGE-2021-3521R1 as invited by the action editor, Nelson Cowan

# Statistically-based chunking of non-adjacent dependencies

Erin S. Isbilen[1,2], Rebecca L.A. Frost[3,4], Padraic Monaghan[5,6], and Morten H. Christiansen[1,2,7]

*Cornell University, Department of Psychology*[1]

*Haskins Laboratories*[2]

*Max Planck Institute for Psycholinguistics, Nijmegen, Language Development Department*[3]

*Edge Hill University, Department of Psychology*[4]

*Lancaster University, Department of Psychology*[5]

*University of Amsterdam, Amsterdam Center for Language and Communication*[6]

*Aarhus University, Interacting Minds Centre and School of Communication and Culture*[7]

**Keywords:** chunking; statistical learning; non-adjacent dependencies; memory; language acquisition

Word count: 10,300

**Please address correspondence to:**
Erin S. Isbilen
Haskins Laboratories
300 George Street, # 900
New Haven, CT 06511
Phone: (203) 865-6163
e-mail: erin.isbilen@yale.edu

**Author note:** We would like to thank Olivia Wang, Eleni Kohilakis, Dante Dahabreh, Jake Kolenda, Phoebe Ilevebare, Farrah Mawani, Emily Zhang, Sophia Zhang, Daniel Mead, Linda Webster, Gauri Binoy, Khalid Mansouri, Kelly Chan and Sharukh Khan for their help collecting and coding the data. We would also like to thank Stewart McCauley for his advice on the linear mixed-effects models presented in this paper. This work was in part supported by the NSF GRFP (#DGE-1650441) and two Cornell Department of Psychology grants awarded to ESI. This work was also supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (UK) [ES/L008955/1]. The preliminary results of Experiment 1 were presented at the Cognitive Science Society's annual conference in July 2018 in Madison, Wisconsin. All materials for the experiments, data, and analysis code are available via the Open Science Framework: https://osf.io/k8mys/.

**Abstract**

How individuals learn complex regularities in the environment and generalize them to new instances is a key question in cognitive science. While previous investigations have advocated the idea that learning and generalizing depend upon separate processes, the same basic learning mechanisms may account for both. In language learning experiments, these mechanisms have typically been studied in isolation of broader cognitive phenomena such as memory, perception, and attention. Here, we show how learning and generalization in language is embedded in these broader theories by testing learners on their ability to chunk non-adjacent dependencies—a key structure in language but a challenge to theories that posit learning through the memorization of structure. In two studies, adult participants were trained and tested on an artificial language containing non-adjacent syllable dependencies, using a novel chunking-based serial recall task involving verbal repetition of target sequences (formed from learned strings) and scrambled foils. Participants recalled significantly more syllables, bigrams, trigrams, and non-adjacent dependencies from sequences conforming to the language's statistics (both learned and generalized sequences). They also encoded and generalized specific non-adjacent chunk information. These results suggest that participants chunk remote dependencies and rapidly generalize this information to novel structures. The results thus provide further support for learning-based approaches to language acquisition, and link statistical learning to broader cognitive mechanisms of memory.

## 1. Introduction

The natural world is awash with statistical regularities, from which individuals can glean the structure of the environment. Yet successful learning entails more than the acquisition of distinct instances from the input: it requires learners to use this information flexibly and extrapolate to novel situations. Investigations into how individuals learn and generalize have a long pedigree in psychology, spanning the domains of episodic learning (Bauer & Dow, 1994), vision and motor control (Poggio & Bizzi, 2004), and language (Wolff, 1982). However, the specific mechanisms and representations involved in these processes, and how individuals move from encoding individual items to forming category-based generalizations, remains an area of debate.

In recent years, one formative memory process has advanced to the frontlines of many discussions on learning across cognitive domains: chunking. Chunking has long been recognized as a foundational cognitive process, enabling the grouping of discrete elements into larger units to alleviate the limits of working memory—a major challenge to the human perceptual system (Cowan, 2001; Miller, 1956). It plays a key role in many higher-level skills such as chess (Chase & Simon, 1973; Gobet & Simon, 1998) and the perception and production of language in real time (Christiansen & Chater, 2016). Chunking has even been implicated as a central component in one of the most powerful means of learning from the regularities present in the environment: statistical learning (e.g., Christiansen, 2019; Perruchet & Pacton, 2006).

Though often discussed as a single mechanism, and investigated in relative isolation from other psychological processes, the phenomenon known as statistical learning may actually comprise a suite of computations, with distinct cognitive processes handling different aspects of learning (Frost, Armstrong, Siegelman & Christiansen, 2015; Frost, Armstrong & Christiansen, 2019). Indeed, mounting evidence highlights the contribution of memory processes, and of

chunking in particular, to statistical learning and the many behaviors it accounts for. At a theoretical level, chunking has been defined as a mechanism by which distributional regularities are used to form discrete representations of an input, especially in the linguistic domain. Numerous chunking-based computational models can approximate human statistical learning of language-related distributional patterns (e.g., PARSER: Perruchet & Vinter, 1998; TRACX: French, Addyman, & Mareschal, 2011), illustrating how chunking can enable the cognitive system to combine co-occurring elements into larger units to represent specific items from an input (e.g., using the frequent co-occurrence of syllables "A" "B" and "C" to form the word "ABC"). Chunking models can also simulate children's natural language acquisition, comprehension and production by leveraging transitional probabilities to define multiword chunks—a finding which extends to numerous languages (McCauley & Christiansen, 2019a).

Behaviorally, chunking-based recall tasks can capture key results in statistical learning, including the landmark study by Saffran, Aslin and Newport (1996), which demonstrated that young infants can rapidly pick up on patterns of syllable co-occurrence in an artificial language consisting of trisyllabic nonsense words. In a recent study, after a brief exposure to a similar artificial language, adult participants recalled syllable sequences that adhered to the statistics of the language significantly better than sequences containing the same syllables in a random order (Isbilen, McCauley, Kidd, & Christiansen, 2020). Participants even recalled specific trigram syllable chunks (or words) from the artificial language, suggesting the involvement of chunking during statistical learning that enabled the representation of whole chunks of information. Similar results were obtained in a follow-up study with 5-6-year-old children (Kidd, Arciuli, Christiansen, Isbilen, Revius, & Smithson, 2020). Just as high-frequency chunks in natural language aid the retention of items in memory (e.g., recalling the letters *"ciafbiusa"* proves easier than recalling

*"uacfisbia,"* as it contains the chunks "CIA," "FBI," and "USA"; Cowan, 2001), the chunking of novel statistical patterns confers comparable memory advantages by reducing cognitive load.

Despite the promise of the chunking account of statistical learning, supporting evidence is still critically limited. Most prior observations of statistically-based chunking focus on the processing of adjacent regularities—that is, relationships between items that immediately follow one another in speech. Yet, language also contains dependencies between elements that do not occur directly next to one another in a sequence, and the learning of these structures necessitates more than rote memorization alone. These non-adjacent (or long-distance) dependencies are ubiquitous in many of the world's languages, allowing for flexible usage (e.g., *is_ing => is sitting, is always talking,* etc.; *un__ed => uncovered, uncensored, unrestrained,* etc.) and linguistic productivity—one of the hallmarks of human language (Hockett, 1959). From the viewpoint of statistical learning, such non-adjacent dependencies constitute reliable statistical relationships between elements that are separated by one or more intervening items (e.g., in the sequence AXC, A and C reliably co-occur but the identity of X varies). They can be learned at the item level (e.g., specific AXC combinations), as well as at the structural level (A-C pairings), along with the ability to generalize these structures to novel instances (e.g., AZC, where Z represents a new item that was not previously encountered in the A-C structure). Non-adjacent dependency learning thus provides a case study for the longstanding debate of how individuals move from encoding specific items to forming generalizations over them (Goldberg, 2006; Radulescu, Wijnen, & Avrutin, 2019), and for determining whether these two abilities rely upon the same suite of statistical computations or require separate processes. It also provides a study of whether chunking is constrained to the grouping of adjacent relations, or if such memory processes are more flexible than previously assumed.

Several studies have successfully demonstrated that adults (e.g., Frost & Monaghan, 2016; Gómez, 2002; Peña, Bonatti, Nespor, & Mehler, 2002; Perruchet, Tyler, Galland, & Peereman, 2004; Romberg & Saffran, 2013), and infants (e.g., Frost et al., 2020; Gómez, 2002; Gómez & Gerken, 1999; Marchetto & Bonatti, 2013; 2015) can acquire non-adjacent dependencies using statistical learning. Indeed, infants as young as six-and-a-half months can leverage redundancy between non-adjacent syllables to boost recognition of trained target syllables (e.g., *ko ba ko*, where *ba* is the target syllable), suggesting that they are sensitive to non-contiguous information starting early in development (Goodsitt, Morse, Ver Hoeve, & Cowan, 1984). However, the precise mechanisms subserving the acquisition of non-adjacent dependencies—and whether they differ from those used to learn adjacent dependencies—have been subject to much discussion, particularly regarding whether these computations draw on statistical learning processes alone or require more complex algebraic operations (see e.g., Frost & Monaghan, 2016; Peña et al., 2002; and Perruchet et al., 2004). While increasing evidence is converging on the notion that non-adjacent dependencies can be discovered via statistical learning (Wilson, Spierings, Ravignani, Mueller, Mintz, Wijnen, Van der Kant, Smith, & Rey, 2020), the extent to which chunking is involved in this process remains highly contested (Endress & Bonatti, 2016).

Non-adjacent dependencies present a formidable challenge to theories that view learning as proceeding through the memorization and recognition of structure. In fact, there is limited evidence that non-adjacent information can be represented in memory as chunks at all, either by human learners or in computational models (e.g., Perruchet & Vinter, 1998; French et al., 2011). By definition, statistical chunking involves the grouping together of elements on the basis of co-occurrence statistics, and it is conceivable that this may extend to non-adjacent relations. However, Kuhn and Dienes (2005) state that "chunking models are very good at learning local dependencies

but cannot learn nonlocal dependencies." Similarly, Bonatti, Peña, Nespor, and Mehler (2006) claim that "much remains to be done before researchers can conclude that humans rely on chunking, as opposed to computing distant transitional probabilities, to capture non-adjacent relations among components of a continuous stream," with both accounts positing the acquisition and representation of rules rather than chunks.

The rule-based framework purports that the learning of words (e.g., AXC) and non-adjacent structure (e.g., A-C pairings) require separate mechanisms. According to this framework, while basic statistical computations are sufficient for acquiring individual AXC items from speech, learning A-C structural relations are thought to require complex, "algebraic" computations involving rule-like representations to enable generalization (Endress & Bonatti, 2007; 2016; Endress Cahill, Block, Watumull, & Hauser, 2009; Endress, Nespor, & Mehler, 2009; Peña et al., 2002). This account further suggests that positional memory mechanisms may be sufficient to explain sensitivity to non-adjacent dependencies (Endress & Bonatti, 2016), with syllables at the edges of these structures being encoded rather than the non-adjacent structure as a whole (Endress & Bonatti, 2007; Endress & Mehler, 2009). By these views, statistically-based chunking is insufficient to account for the behavioral data, and chunked representations are rejected in favor of rules and the memorization of ordinal position. These theories parallel classical linguistic frameworks, which posit that statistical computations are insufficient for language acquisition (Chomsky, 1957; 1980), which must instead rely on symbolic grammatical inference to generalize beyond the limited exemplars in the input.

The question of how non-adjacent structures are represented has been the focus of much attention. Understanding these representations is of particular importance, as it is from the data of studies targeted to probe these representations that researchers often infer the nature of the

computations employed during learning and generalization. Findings from the chunking literature suggest a possible role for chunking in learning non-local structures—a finding that would be expected if such memory processes are indeed integral to learning at large. For instance, various studies of artificial grammar learning (AGL; see Perruchet, 2019, for a review), which have strong parallels with statistical learning approaches, have demonstrated the central role of chunk strength (the relative frequency with which bigrams or trigrams in a test item occurred together during training; Knowlton & Squire, 1994; 1996) in the learning of simple, variable grammars while discounting the acquisition of rules (Kinder & Assmann, 2000). Indeed, chunking models appear well suited to capturing human statistical learning representations in the linguistic domain (e.g., Giroux & Rey, 2009; Frank, Goldwater, Tenenbaum, 2010), and in the spatial domain (Orbán, Fiser, Aslin, & Lengyel, 2008), where information is not necessarily contiguous.

Importantly, the chunking perspective makes specific predictions about both the computations and representations involved in statistical learning. Based on statistical regularities, the chunking process combines recurring items into larger units online during learning. This leads to the formation of concrete, chunked representations of the input, such as words, phrases, or multiword units that can be used to formulate novel constructions (e.g., McCauley & Christiansen, 2019a). These ideas relate to usage-based frameworks within the study of language acquisition (e.g., Goldberg, 2006; Lieven, Pine, & Baldwin, 1997; Tomasello, 2003), where learners are thought to acquire specific items from the input through experience, which serve as the foundation for generalization and productivity. Unlike classical linguistic theories, such exemplar-based learning is not language-specific, but is thought to apply across cognitive domains.

To reappraise the relationship between acquisition and generalization, and the role of memory therein, the current paper employs the statistically-induced chunking recall task (SICR;

Isbilen et al., 2020), using non-adjacent dependency learning as a test case. SICR presents both statistically legal items (composed of two trisyllabic target words) and illegal strings (the same syllables randomized) from a trained artificial language, which participants are asked to recall out loud in the correct order. If participants have chunked the input language into individual words during training (e.g., *"abcdef"* => "abc" "def"), then recall of the trained items should yield significantly higher accuracy than recall of the random strings (e.g., "*efbdca*"). Furthermore, as the SICR data is transcribed then scored syllable-by-syllable, it is possible to directly examine participants' representations in a manner that standard forced-choice tasks do not afford. The more detailed memory-based measures of SICR thus provide specific insights into the representations formed during processing. Participants' productions can be analyzed for specific chunk formation—recall of full words from the trained target strings, and in the present case, recall of full non-adjacent dependencies. Sensitivity to statistical structure can thus be measured by comparing recall of the target strings to recall of the foils, which serve as a baseline working memory measure.

The SICR paradigm is modelled on key findings from the memory literature, indicating that immediate recall abilities are fundamentally shaped by long-term distributional learning. For instance, nonwords that adhere to the phonotactic patterns that occur regularly in natural speech are recalled more accurately than those based on infrequent phoneme sequences (Gathercole, Frankish, Pickering, & Peaker, 1999), and high-frequency digit combinations are recalled more accurately than lower-probability numerical sequences (Jones & Macken, 2015; see Cowan, Rouder, Blume, & Saults, 2012 for evidence that memory for high-frequency chunks of linguistic items also benefit from a similar advantage). Just as chunking-based recall tasks such as nonword repetition and serial recall provide key insights into both children's processing abilities and their

current degree of real-world linguistic knowledge (Jones, 2012), SICR captures the same effects with the learning of artificial languages.

The current research tests the chunking account of learning, seeking to determine whether general memory processes can capture the acquisition of non-adjacent structures in a statistical learning paradigm and whether they are represented as chunks. Crucially, we assess whether chunking can extend beyond the mere recognition of non-adjacent relationships to items that contain generalizations of these structures. The first experiment investigates the acquisition and generalization of non-adjacent dependencies, following the methods of Frost and Monaghan (2016). The second experiment addresses the long-standing debate concerning the precise nature of the representations formed during non-adjacency learning, and tests whether such structures are represented as chunks or whether participants simply encode the relative positions of individual syllables. Together, these two experiments speak to the general issue of how non-adjacent structures are learned and generalized, the role of memory-based chunking therein, and their relation to language and cognition.

We hypothesized that non-adjacent dependency learning would boost SICR performance, suggesting that the chunking of accrued linguistic distributional information can facilitate short-term recall for sequences comprising non-adjacent dependencies in the same manner that has been observed for adjacent dependencies (e.g., Chen & Cowan, 2009; Jones & Macken, 2015). Furthermore, we hypothesized that participants would encode specific non-adjacent pairings (rather than encoding the relative positions of syllables alone or adhering to the rule structure of the stimuli). We predicted that these representations would also facilitate generalization, suggesting that statistical learning and generalization may rely upon the same memory-based mechanisms.

## 2. Experiment 1: Statistically-based chunking of non-adjacent dependencies

The first experiment investigates whether general memory processes can support the learning and generalization of non-adjacent structures in an artificial language. To test this, we utilized the SICR task (Isbilen et al., 2020): a recall task where participants reproduce strings of syllables that either cohere with or violate the statistics of the presented language. Because this task entails the production of test strings, participants' responses can be transcribed and analyzed for the presence (or absence) of the specific structures from the training input. This task thereby provides a direct window into participants' representations of the language, and the potential involvement of chunking during learning. For instance, we can examine the data for two key signatures of statistical chunking behavior: whether participants recall full trigrams (i.e., full words) and the number of non-adjacent pairs recalled, with the generalization trials as the ultimate test of non-adjacent chunking. If participants recall specific trained non-adjacent pairs on the novel trials, this suggests that such pairs are represented as a single chunk in memory. We also administered a two-alternative forced-choice task (2AFC) to provide an additional measure of learning, and to make contact with the existing body of literature that tests non-adjacency learning using 2AFC (e.g., Endress & Mehler, 2009; Frost & Monaghan, 2016; Peña et al., 2002; Perruchet & Poulin-Charronnat, 2012).

We predicted that learning and generalization would be exhibited on both tests, with recall of the target items being superior to foils on SICR, and with 2AFC performance being greater than chance. Additionally, we hypothesized that participants would recall significantly more legal trigrams (or full words) on the word learning trials, and that the statistical facilitation from legal non-adjacent dependencies would also lead to improved recall of trigrams on the generalization

trials. Lastly, we hypothesized that participants would recall specific non-adjacent chunks from the grammatical items, suggesting that they can in fact chunk such structures.

## 2.1. Method

### *2.2. Participants*

Forty-nine undergraduates (30 females, 19 males; age: *M*= 19.43, *SD*= 1.30) from Cornell University were recruited. All participants were native speakers of American English, with no known language or hearing disorders. Participation was compensated with course extra credit.

### *2.3. Materials*

This study utilized the same artificial language as Frost and Monaghan (2016), which was adapted from Peña et al. (2002). The language comprised 9 syllables, which were used to create three non-adjacent dependencies (e.g., A-C pairings), which each featured three different middle syllables (e.g., X syllables in AXC), yielding nine distinct tri-syllabic words that were heard during training ($A_1X_1C_1$, $A_1X_2C_1$, $A_1X_3C_1$; $A_2X_1C_2$, $A_2X_2C_2$, $A_2X_3C_2$; $A_3X_1C_3$, $A_3X_2C_3$, $A_3X_3C_3$). Plosives (*be, du, ga, ki, pu, ta*) were used for the first and third syllables of each non-adjacent dependency, whereas continuants (*fo, li, ra*) were used for the middle syllables (e.g., *dufoki, duliki, duraki; gafobe, galibe, garabe; tafopu, talipu, tarapu*).[1] To ensure that the study's results were not due to the particular features of a single artificial language, four different languages with different A-C pairings were created, and were counterbalanced across participants. The words in each language

[1] This is in line with the language by Peña et al. (2002). Phonological similarity between syllables supports the acquisition of non-adjacent dependencies (Newport & Aslin, 2004; but see Frost, Isbilen, Christiansen & Monaghan, 2019; and Onnis, Monaghan, Richmond, & Chater, 2005, for evidence that this is not essential for learning).

version were concatenated together into a single auditory file that was presented during the training phase, with a five second fade in and fade out to prevent participants from using the onset and offset of the file as a cue for determining word boundaries. The transcriptions of the words for each language version can be found in Appendix 1a.

A further nine generalization words that were not present during training were created to test how well participants could generalize their knowledge of the trained non-adjacent dependencies. These generalization words were composed of the same non-adjacent dependencies as the input words but featured novel intervening syllables not heard during training (e.g., Z syllables in AZC). Like the input words, continuants were used for the middle syllables (*thi, ve, zo*). The transcriptions of the generalization words for each language version can be found in Appendix 1b.

For SICR, 26 six-syllable-long strings were created, in line with those used by Isbilen et al., (2020). Of these, nine were composed of two concatenated words (e.g., $A_1X_3C_1A_2X_1C_2$), nine were composed of two concatenated generalization words (e.g., $A_1Z_3C_1A_2Z_1C_2$), and eight were foils (to maintain equal numbers of each item type: four word-learning and four generalization foils). The foils used the same syllables as the target items but in a scrambled order, avoiding both the adjacent and non-adjacent regularities of the artificial language. These foils served as a baseline working memory measure, which performance on the target items was compared against to gauge learning. All SICR test items can be found in Appendix 2a.

For 2AFC, the same eighteen foil words from Frost and Monaghan (2016) were used: nine part-word foils, and nine generalization foils for each language version. The part-word foils spanned word boundaries (e.g., $X_1C_1A_2$, $C_2A_1X_1$), and were presented alongside the input words to test how well participants had picked up on the language's underlying structure. Similarly, the

generalization foils were also part-words that spanned word boundaries, but with the X syllables replaced with novel, unheard syllables (e.g., $Z_1C_1A_2$, $C_2A_1Z_1$), which were presented at test with the target generalization words. All 2AFC foil words can be found in Appendix 2b.

All stimuli were synthesized using the Festival speech synthesizer (Black, Taylor, & Caley, 1990), with each individual tri-syllabic sequence lasting approximately 700ms. Both stimulus presentation and data collection utilized E-prime 2.0 (Schneider, Eschman, & Zuccolotto, 2002). The study was approved by Cornell University's Institutional Review Board, and participants signed a consent form prior to participation. The stimuli are available on OSF: https://osf.io/k8mys/.

*2.4. Procedure*

First, participants were trained on the artificial language. The nine words were randomized to produce a continuous stream, and participants were instructed to listen to the language carefully and pay attention to the words it might contain. Each word was presented 100 times (with each non-adjacency pair occurring 300 times), and training lasted approximately 10.5 minutes. The nine words were randomized and concatenated such that there was no immediate repetition of individual AXC words.

Following exposure, participants completed two tasks: SICR and 2AFC. The order of these tasks was counterbalanced across participants. For SICR, the twenty-six strings described above were presented for recall: nine that tested word acquisition, nine that tested generalization, and eight random foil strings, to maintain equal numbers of word learning and generalization foils (four each). Participants were told that they would be assessed on how well they could reproduce the syllables present in the artificial language. They were then asked to listen to each string

carefully, and to repeat the entire string out loud in the correct order as accurately as possible into a microphone as soon as it was finished playing. The order of the SICR items was randomized across individuals, and participants were not informed of any underlying structure present in the strings.

For 2AFC, participants heard eighteen pairs of words: one target word, and one foil word per trial. Of these, nine pairs tested acquisition of the input words, and nine tested generalization. For this task, participants were asked to listen to each word pair carefully and report which of the two items best matched the artificial language they were trained on. The order of all 2AFC trials was randomized across individuals.

### *2.5. Results*

#### *2.5.1. SICR results*

All data and R scripts are available on OSF: https://osf.io/k8mys/. Prior to analysis, participants' verbal responses on the SICR task were transcribed by two coders who were naive to the purpose of the study and its design (see Isbilen et al., 2020, for an in-depth guide on the transcription of SICR sequences). In line with Isbilen et al. (2020) and methods used in the nonword repetition literature (Botnivick & Bylsma, 2005), consistent syllable mispronunciations (e.g., a participant routinely says "le" for the target syllable "li") were transcribed as correct (i.e., as "li" rather than "le"), as such mispronunciations indicate differences in how participants perceive the syllables produced by the speech synthesizer. In addition, an anchoring procedure was used to align participants' productions as closely as possible to the target items, identical to what is done for many nonword repetition tasks (e.g., Dollaghan & Cambell, 1998; Weismer, Tomblin, Zhang, Buckwalter, Chynoweth, & Jones, 2000). For instance, if a target stimulus was "taragabeliki" and

the participant produced "taraki," this would be transcribed as "tara------ki" (with a dash denoting each missed letter). This ensured that participants were granted credit for all of the syllables that were correctly recalled, even if they returned fewer syllables than they were presented with. In cases where there were false starts (e.g., the participant started producing a syllable, paused, then started again) and self-corrections (where a participant corrected their production of a syllable or item), the original production was ignored in favor of the second/corrected production. Non-responses on a trial were automatically given a score of zero.

Following transcription, participants' responses on both the word learning and generalization trials were scored for accuracy using four different measures. The first measure was total accuracy (the total number of syllables that participants recalled in the correct order), which allowed us to gauge how well participants performed on each string as a whole, and whether statistical learning conferred a general memory advantage. The second measure was bigram accuracy (the total number of adjacent two-syllable combinations recalled within words, out of four possible pairs: e.g., $A_1X_1$, $X_1C_1$, $A_2X_2$, $X_2C_2$ in the target sequence $A_1X_1C_1A_2X_2C_2$), which indicates how well participants encoded the adjacent bigram information within the presented structures. The third measure was trigram accuracy (the total number of trigrams or full words correctly recalled in each string, out of two possible pairs: e.g., $A_1X_1C_1$ & $A_2X_2C_2$), which revealed whether participants chunk entire words in the target word learning trials, and whether non-adjacency learning enabled better retention of novel words in the generalization trials. The final measure was non-adjacent dependency accuracy. For the target sequences, this was the number of A-C pairings that participants recalled from each string, out of two possible pairs: e.g., $A_1\_C_1$ & $A_2\_C_2$. For the foils, this score was calculated based on recall accuracy for pairs of syllables in the analogous positions, i.e., syllables 1 & 3, and syllables 4 & 6. This measure allowed us to

determine whether participants chunked specific non-adjacent syllable combinations in the target items.

The target item scores were then compared against those of the foil items, to test statistical learning against baseline working memory. We report the results from linear mixed effect model analyses below, which used the "lmerTest" package (Kuznetsova, Brockhoff, & Christensen, 2017) in the statistical software R, version 4.0.2 (R Core Team, 2020). For the linear mixed effects models, the models were built incrementally, and subjects and test items were included as random effects, and word type (target vs. foil) as a fixed effect. Language version was not included as a separate random effect in the analyses, as it was already redundantly coded in the test item variable. Each SICR measure (total, bigram, trigram and non-adjacent dependency accuracy) was modeled separately.

On the word learning trials, when controlling for subjects and test items, the fixed effect of word type on total accuracy was highly significant (model improvement over model containing only random effects: $\chi^2(1) = 43.16$, $p<.0001$), with participants correctly recalling significantly more syllables in the target items than in the foil items (difference estimate= -1.16, *SE*=.14, *z*= -8.15, $p<.0001$). There was also a significant effect of word type on bigram recall (model improvement over model containing only random effects: $\chi^2(1) = 36.87$, $p<.0001$; difference estimate= -.87, *SE*=.12, *z*= -7.41, $p<.0001$) and on trigram recall (model improvement over model containing only random effects: $\chi^2(1) = 36.73$, $p<.0001$; difference estimate= -.42, *SE*= .06, *z*= -7.25, $p<.0001$), with significantly higher recall accuracy of legal adjacent syllable combinations and full trigram words encountered in the input. Finally, there was a robust effect of word type on non-adjacent dependency accuracy (model improvement over model containing only random effects: $\chi^2(1) = 34.08$, $p<.0001$), with significantly more non-adjacent dependencies (syllables in

the first & third serial positions and the fourth & six serial positions) being accurately recalled in the target items than the foils (which contained no non-adjacent structure; difference estimate= -.43, *SE*= .06, *z*= -6.92, *p*<.0001).

For the generalization items, there was a significant effect of word type on SICR total accuracy (model improvement over model containing only random effects: $\chi^2(1) = 25.77$, *p*<.0001), with participants recalling significantly more syllables in the target strings (difference estimate= -1.00, *SE*= .17, *z*= -5.78, *p*<.0001). The same pattern was observed for bigram accuracy (model improvement over model containing only random effects: $\chi^2(1) = 8.72$, *p*= .003; difference estimate= -.52, *SE*= .17, *z*= -3.08, *p*= .004) and trigram accuracy (model improvement over model containing only random effects: $\chi^2(1) = 36.73$, *p*<.0001; difference estimate= -.42, *SE*= .06, *z*= -7.25, *p*<.0001), with better accuracy for legal adjacent syllable combinations and full novel trigram words within the target sequences. Lastly, a strong effect of word type was also observed for non-adjacent dependency accuracy (model improvement over model containing only random effects: $\chi^2(1)= 21.95$, *p*<.0001), with participants recalling significantly more non-adjacent pairs within the target than foil items (difference estimate= -.37, *SE*= .07, *z*= -5.20, *p*<.0001).

There was no significant difference between word learning and generalization as measured by the SICR difference scores (target item score minus foil item scores), for either total accuracy, or non-adjacent dependency accuracy (both *p*=.10). However, a significant difference was observed on bigram recall (*t*(48)= 4.12, *p*=.0002, *d*= .59) and trigram recall (*t*(48)= 3.86, *p*=.0003, *d*= .55). Participants recalled the bigrams and trigrams they were exposed to significantly better than the items with novel middle syllables. Mean performance on these different measures (i.e., the average proportion of syllables, bigrams, trigrams and non-adjacent dependencies recalled

across all participants for each trial type) can be found in Table 1, and the serial position curves in Figure 1.

Table 1

*Summary statistics for SICR by item type (proportion correct)*

*Word learning*

| | Syllables | | | Bigrams | | | Trigrams | | | NADs | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| Target | .22 | .07 | .06 – .33 | .17 | .08 | 0 – .33 | .14 | .09 | 0 – .31 | .16 | .09 | .02 – .31 |
| Foil | .09 | .04 | .02 – .17 | .05 | .04 | 0 – .15 | .04 | .04 | 0 – .13 | .05 | .05 | 0 – .15 |

*Generalization*

| | Syllables | | | Bigrams | | | Trigrams | | | NADs | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| Target | .18 | .07 | .05 – .33 | .11 | .08 | 0 – .32 | .09 | .08 | 0 – .31 | .13 | .09 | .02 – .33 |
| Foil | .05 | .03 | .01 – .12 | .03 | .03 | 0 – .10 | .02 | .03 | 0 – .10 | .03 | .03 | 0 – .10 |

**[[Insert Figure 1 here]]**

*2.5.2. 2AFC results*

As predicted, participants performed significantly above chance on both the word learning and generalization 2AFC trials (Word learning: $t(48)= 18.44$, $p<.0001$, $d= 2.63$; Generalization: $t(48)= 6.61$, $p<.0001$, $d=.94$). Participants' accuracy was significantly higher on the word learning than on the generalization trials, $t(48)=5.77$, $p<.0001$, $d= .82$. The mean scores for each trial type can be found in Table 2.

Table 2

*Summary statistics for 2AFC by item type (proportion correct)*

| | Mean | SD | Range |
|---|---|---|---|
| Word learning | .84 | .13 | .56 – 1 |
| Generalization | .70 | .21 | .22 – 1 |

*2.6. Discussion*

Experiment 1 provides strong evidence that participants could identify individual words in an artificial language and detect the non-adjacent dependencies it contained. Since both words and structure were defined by (adjacent and non-adjacent) transitional statistics, these data lend further support to the notion that human learners can perform computations over the distributional regularities contained in speech, even for dependencies that occur over a distance. Importantly, participants also successfully generalized statistically-learned chunks of non-adjacent information, and could do so in relative synchrony with learning the precise sequences that occurred, rather than requiring the progressive formation of rule-like representations after item learning had been mastered.

Importantly, the results also indicate that our ability to perform these tasks may be driven (at least in part) by statistically-based chunking. On SICR, recall was significantly better for the structured strings than random foils, with participants recalling significantly more bigrams, trigrams and non-adjacent pairs for structured sequences—structures that have historically posed a challenge for chunking models. This mirrors prior work demonstrating that individuals in serial recall tasks can remember items that are interpolated by additional information, which may be seen as a kind of non-adjacent structure (Baddeley, Papagno, & Andrade, 1993). For instance, when tasked with recalling strings of numbers interleaved by words (e.g., 7-wit-9-bond-6), participants can successfully ignore the words and recall the numbers, suggesting that distal information can be held in verbal working memory. This effect replicates for visual-spatial serial information (Nicholls, Parmentier, Jones, & Tremblay, 2007), suggesting that it is a general property of memory across domains. Recall in these studies was somewhat lower for interpolated stimuli than when recall items were contiguous (with similar findings by Greene, Elliot, & Smith, 1988; Hitch, 1975; Murray, 1966), although this effect was small in both Baddeley et al. (1993) and Nicholls et al. (2007). The current study moves beyond these results, investigating memory for non-adjacent information following a training phase on an artificial language, to measure learning-induced changes to recall. While previous studies have shown that participants can suppress interleaved information, here we find that individuals can recall adjacent and non-adjacent information simultaneously and with comparable accuracy, facilitated by statistical learning.

The enhanced trigram recall on the target word learning trials suggests that participants had chunked the syllables into wholistic word-like representations during training. Higher trigram recall on the target generalization items implies that the chunking of non-adjacent dependencies facilitated recall of the new intervening items by reducing memory load. These findings with non-

adjacent patterns parallel prior research underscoring the contribution of chunking to representing learned adjacent items in memory in serial recall (Chen & Cowan, 2009) and non-word repetition (Jones, Gobet, Freudenthal, Watson, & Pine, 2014), and how long-term statistical knowledge interacts with these abilities (Jones, 2012; Jones & Macken, 2015; Cowan et al, 2012). Just as prior studies report memory advantages for learned chunks of adjacent information, here we extend these results to the statistical learning of non-local structures, demonstrating how the statistical learning of such structures are relevant for broader theories of cognition and memory.

Additionally, participants also recalled significantly more trained non-adjacent dependencies—specific A-C combinations—in the target items than the foils (which did not contain this structure). This was true for both the word learning and generalization sequences, suggesting that participants formed chunked representations of these dependencies, which they could use flexibly in novel instances. Participants' knowledge of the non-adjacent dependencies therefore appears to facilitate their ability to recall information on the generalization trials, while the word learning trials demonstrate evidence of specific item learning. To our knowledge, the present study is the first to demonstrate that long-term distributional knowledge facilitates memory for non-adjacent items in the same manner that has been observed for adjacent items.

Using two assessments of learning (SICR and 2AFC), we replicated Frost and Monaghan's (2016) finding that adults can identify words and word-internal dependency structures together during learning, from statistical information alone (without the need for additional cues, e.g., Fló, 2021; Peña et al., 2002), indicating that these tasks may be underpinned by similar statistical learning and memory processes (see Frost & Monaghan, 2016, for further discussion). Our results thus provide further evidence for the notion that learning and generalization may simply be different outcomes of the same statistical learning processes (Aslin & Newport, 2012) rather than

generalization requiring separate rule-like computations, suggesting a more unified framework for non-adjacent dependency learning and language acquisition at large. However, some outstanding questions about how non-adjacent information is represented in memory remain, including whether such information is represented as chunks or merely encoded positionally.

## 3. Experiment 2: Chunked representations of non-adjacent dependencies

Experiment 1 provided initial support for the chunking of whole words, and the specific non-adjacent syllables within them. However, an alternative possibility may explain the results: that participants merely encoded the relative positions of the syllables rather than the non-adjacent dependencies as chunks (Endress & Mehler, 2009). Possible evidence against the chunking hypothesis and for the positional information hypothesis comes from participants' inability to distinguish trained non-adjacent dependencies (e.g., $A_1XC_1$) from items that violate the language's chunk information while preserving its positional information, by replacing the final syllable of one trained dependency with the final syllable of another (e.g., $A_1XC_2$). These items, first introduced as "class-words" (Endress & Bonatti, 2007), and later as "phantom words" (Endress & Mehler, 2009), reportedly led participants to have false memories of hearing these words during training, thus suggesting that positional information—and not chunk information—is the outcome of non-adjacent dependency learning. However, others have since shown that participants exhibit a general preference for trained words over phantom words—consistent with the predictions of chunking-based computational models (Perruchet & Poulin-Charronnat, 2012). However, what participants specifically represent—trigrams, non-adjacent dependencies, or both—remains an open question.

Given that most prior investigations into this phenomenon have utilized 2AFC, we sought to revisit the phantom word effect using SICR. As SICR affords more specific insights into learners' representations, we reasoned that it may grant clearer insights into whether participants acquire chunk information or only positional information. Unlike previous work on this topic, we tested both acquisition and generalization rather than acquisition alone. We hypothesized that participants would differentiate between trained and phantom items, showing learning of specific chunks rather than solely positional information. However, we hypothesized that this would only be on the SICR task, due to the increased specificity that the recall data provides about participants' representations relative to 2AFC. We pre-registered our predictions for the experiment prior to data collection (https://aspredicted.org/7z7m8.pdf).

### *3.2. Participants*

We collected data from 75 participants (49 females, 25 males, 1 non-binary; age: $M$= 20.65, $SD$= 3.07). Due to the COVID-19 pandemic, this experiment was conducted online. We recruited from both the Cornell University undergraduate population (N= 50) and from the Prolific participant recruitment platform (N= 25). Participation from the Prolific subject pool was limited to university students who were native speakers of American English, to maintain comparability between the two samples. There were no significant differences in performance between the two samples on either SICR or 2AFC ($p$= .24 or greater). Therefore, the data from both samples were combined and analyzed together.

This pre-registered sample size was determined by a power analysis based on the results of a pilot experiment run in-lab, wherein we observed a SICR effect size of approximately $d$=.41.

Due to the online format of the current study, we anticipated that the observed effect sizes might be slightly smaller than the pilot study that was conducted in-person prior to lab closures (e.g., due to potential delays between stimulus presentation and when participants are cued to start repeating back stimuli because of differences in internet connectivity, differences in headphone types across participants, slight differences in the volume at which participants listen to stimuli, etc.). We therefore conducted a power analysis based on a reduced effect size of $d$ = .21 (half of the effect size observed in the pilot study). All participants were native speakers of American English, with no known history of language or auditory disorders. Participation was compensated with course extra credit or monetary payment.

### *3.3. Materials*

The same four artificial languages as Experiment 1 were used, featuring nine AXC words that were heard during training. In addition, the same 18 SICR and 18 2AFC target items as Experiment 1 were used at test (9 that tested word learning and 9 that tested generalization in each task). For the foils, eighteen phantom words were created. Phantom words were constructed by taking the first syllable of one non-adjacent dependency and pairing it with the final syllable of a different non-adjacent dependency (e.g., $A_1X_1C_2$), to preserve the items' positional information but disrupt their chunk information (i.e., $A_1$ and $C_2$ occur as the first and last syllable in a trisyllabic sequence, but never occurred together within the same trisyllabic word during training). The same method was used to create 4 SICR phantom word learning trials ($A_1X_3C_2A_2X_1C_1$), and 4 SICR phantom generalization trials ($A_1Z_3C_2A_2Z_1C_1$), yielding 8 SICR phantom word trials in total. The bigram information in the foils was carefully balanced, to make sure that no single bigram appeared in the

phantom word strings more than once, and all phantom A-C pairings occurred an approximately equal number of times. The phantom word items for both tasks can be found in Appendix 3.

As in Experiment 1, all new stimuli were generated using the Festival speech synthesizer (Black et al., 1990), using the same voice as the items in the first experiment, with each tri-syllabic string lasting approximately 700 milliseconds. Both stimulus presentation and 2AFC data collection utilized Qualtrics survey software. Participants' spoken responses on the SICR task were recorded using the Zoom conferencing software, in a completely anonymized meeting session (participants logged in with their randomized participant numbers and with no video, to ensure that the data was completely de-identified), as was approved by Cornell University's Institutional Review Board. Participants signed a consent form prior to participation. All stimuli are available on OSF: https://osf.io/k8mys/.

### *3.4. Procedure*

Identical to Experiment 1, participants were first trained on the artificial language for approximately 10.5 minutes. During this time, they were asked to listen carefully to the language and pay attention to any words it might contain. Each word was presented 100 times (and so each non-adjacent dependency was presented 300 times) throughout the course of training.

Following exposure, word learning and generalization were measured using both SICR and 2AFC. The order of these two tests was counterbalanced across participants. In SICR, participants heard strings of syllables over headphones, and were asked to repeat the entire string out loud to the best of their ability. Participants were not informed of the strings' underlying structure. For 2AFC, participants heard stimulus pairs consisting of words and phantom words. On the word learning trials, input words were always paired with phantom input words. On the generalization

trials, generalization words were always paired with phantom generalization words. Participants were instructed to indicate which of the two words best matched the artificial language they were exposed to.

### *3.5. Results*

#### *3.5.1. SICR results*

All data and R scripts are available on OSF: https://osf.io/k8mys/. As in Experiment 1, two coders who were blind to the purpose of the study and its design transcribed the SICR data, using the same procedures described in Section 2.5.1. Participants' productions were then scored for total syllable, bigram, trigram, and non-adjacent dependency accuracy, and learning was measured by comparing responses on the target items to those on the phantom word items. Word learning and generalization responses were modeled separately using linear mixed effects models, with subject and test item as random effects, and word type (target vs. phantom word) as a fixed effect. As in Experiment 1, language version was not included in the models as a separate random effect, as it was already redundantly coded within the test item variable.

When considering the total recall accuracy on the word learning trials (i.e., the total number of syllables recalled), the fixed effect of word type was not significant (model improvement over model containing only random effects: $\chi^2(1)= 2.60$, $p= .11$), with no reliable difference between the number of syllables recalled for target and phantom word strings (difference estimate= -.23, *SE*=.15, *z*= -1.61, *p*= .11). Similarly, there was no significant effect of word type on bigram recall (model improvement over model containing only random effects: $\chi^2(1)= 2.42$, $p= .12$), with performance on both target and phantom strings being approximately equal (difference estimate= -.22, *SE*= .14, *z*= -1.55, *p*= .13). However, there was a significant effect of trigram accuracy (model

improvement over model containing only random effects: $\chi^2(1)= 4.35$, $p= .037$): participants recalled significantly more legal trigrams than phantom trigrams (difference estimate= -.16, *SE*= .08, *z*= -2.11, *p*= .04). Similarly, non-adjacent dependency accuracy was significantly higher for the target items (model improvement over model containing only random effects: $\chi^2(1) = 7.53$, $p= .006$), with participants recalling more non-adjacent dependencies that they were exposed to during training than phantom dependencies (difference estimate= -.18, *SE*= .06, *z*= -2.82, *p*= .007).

For SICR generalization, the data reveal the same pattern of results as those observed for the word learning trials. The linear mixed effects models revealed no significant effect of word type on total accuracy (model improvement over model containing only random effects: $\chi^2(1)= 2.23$, $p= .14$), with participants performing similarly on both item types (difference estimate= -.20, *SE*= .14, *z*= -1.50, *p*= .14). Similarly, there was no significant effect of item type on bigram accuracy (model improvement over model containing only random effects: $\chi^2(1) = 1.25$, $p= .26$; difference estimate= -.13, *SE*= .12, *z*= -1.11, *p*= .27). However, for trigram recall, word type was significant (model improvement over model containing only random effects: $\chi^2(1) = 4.35$, $p= .037$), with participants recalling significantly more trained trigrams over phantom trigrams (difference estimate= -.16, *SE*= .08, *z*= -2.11, *p*= .041). Non-adjacent-dependency recall was also significantly impacted by word type (model improvement over model containing only random effects: $\chi^2(1) = 5.00$, $p=.026$), with participants recalling significantly more non-adjacent dependencies for the target items than in the phantom word items (difference estimate= -.13, *SE*= .06, *z*= -2.62, *p*= .03).

Performance was significantly higher on the word learning trials than on the generalization trials for all SICR measures (total accuracy: $t(74)= 2.63$, $p= .01$, $d= .30$; bigram accuracy: $t(74)= 3.81$, $p= .0003$, $d= .44$; trigram accuracy: $t(74)= 3.63$, $p= .0005$, $d= .42$; non-adjacent dependency

accuracy: $t(74)= 2.04$, $p= .05$, $d= .24$). Mean performance on all SICR measures (e.g., the average proportion of syllables, trigrams or non-adjacent dependencies recalled across all participants for each trial type) can be found in Table 3, and the serial position curves in Figure 2.

Table 3

*Summary statistics for SICR by item type (proportion correct)*

*Word learning*

| | Syllables | | | Bigrams | | | Trigrams | | | NADs | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| Target | .25 | .05 | .13 – .31 | .23 | .06 | .08 – .32 | .21 | .07 | 06 – .31 | .22 | .06 | .06 – .31 |
| Phantom | .12 | .03 | .03 – .15 | .10 | .04 | .02 – .16 | .09 | .04 | .02 – .15 | .10 | .04 | .02 – .15 |

*Generalization*

| | Syllables | | | Bigrams | | | Trigrams | | | NADs | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| Target | .22 | .06 | .08 – .31 | .18 | .07 | .04 – .32 | .16 | .08 | .02 – .31 | .20 | .07 | .06 – .31 |
| Phantom | .11 | .03 | .04 – .15 | .08 | .04 | 0 – .16 | .07 | .04 | 0 – .15 | .09 | .04 | .02 – .15 |

**[[Insert Figure 2 here]]**

### *3.5.1. 2AFC results*

Contrary to our pre-registered predictions, 2AFC word identification performance was significantly above chance, with participants preferring words over phantom words ($t(74)= 5.87$, $p<.0001$, $d=.68$). Generalization performance was also significantly above chance ($t(74)= 1.67$, $p= .05$, $d= .19$), with participants correctly selecting novel but structurally consistent words constructed of target non-adjacent dependencies over phantom dependencies. There was a significant difference between 2AFC word learning and generalization performance, $t(74)= 3.03$, $p= .003$, $d= .35$, with participants performing significantly better on the word learning than the generalization trials. The mean accuracy of each 2AFC measure is reported in Table 4.

Table 4

*Summary statistics for 2AFC by item type*

| | Mean | SD | Range |
|---|---|---|---|
| Word | .61 | .16 | .33 – 1 |
| Generalization | .53 | .17 | 0 – 1 |

### *3.6. Discussion*

In Experiment 2, we expanded on the results of Experiment 1 by performing a closer examination of the representations that learners form during statistical non-adjacent dependency learning. In doing so, we revisited the positional hypothesis proposed by Endress and Mehler (2009), but with two key differences from previous studies. First, we tested generalization in addition to acquisition, to determine whether the effect reported for phantom input words would extend to generalization. Second, we employed SICR to gain more detailed insights into participants' representations for

the newly acquired words/structures, to disentangle whether these comprise chunks or merely positional information.

Our results reveal that participants may become sensitive to multiple kinds of regularities in non-adjacency tasks, including both adjacent and non-adjacent information. However, it also appears that they do encode specific non-adjacent chunks above and beyond the positions of syllables alone. For SICR, contrary to our predictions, no difference was found between recall of the target versus phantom words in terms of the total number of syllables recalled or the number of bigrams recalled. This finding makes sense when considering the overall statistical structure of the target and phantom items: in both cases, the adjacent bigram information was identical. The only differences between the two lay in the trigram and non-adjacency statistics. We observed a robust difference in both the number of legal trigrams and legal non-adjacent dependencies recalled, both on the word learning and generalization trials. Although the distinctions between the target and phantom word items were very subtle in terms of their statistical structure, participants nonetheless recalled significantly more trigrams and non-adjacent dependencies that followed the artificial language's chunk information. These findings dovetail with experimental data on visual statistical learning from Slone and Johnson (2015), demonstrating that participants can distinguish trained triplets from statistically matched illusory triplets, and thus represent chunks rather than statistics. Comparable results have also been reported for linguistic statistical learning (Perruchet & Poulin-Charronnat, 2012; Wang, Zevin, & Mintz, 2019). However, while these previous studies have demonstrated a general preference for trained over phantom items, here we elucidate the specific representations that learners accrue.

The stronger facilitation from legal non-local dependencies has several implications for the kind of information that participants glean from learning. Work from a recent study of online visual

statistical learning shows that there are strong individual differences in the kinds of dependencies—local versus non-local—that participants rely on (Siegelman, Bogaerts, Armstrong, & Frost, 2019), with some individuals preferring one over the other while some attend to both. Overall, adjacent information appears to be easier for participants to process and learn (Trotter, Monaghan, Beckers, & Christiansen, 2020), with participants potentially favoring adjacent over non-adjacent dependencies when such information is present (Gómez, 2002). Indeed, Gómez (2002) found that individuals could only learn non-local over local information when there was ample variability of the middle items in the input—even with the inclusion of additional pause and lexical cues to the non-adjacent structure. The languages in the present experiment possessed no additional cues and very few middle items: only three as opposed to the twenty-four middle items required in Gómez (2002) for participants to endorse grammatical over ungrammatical non-adjacent sequences above chance in a grammaticality judgement task. Here, we contribute further evidence that individuals can in fact learn and represent adjacent transitional probability information and non-adjacent information simultaneously, supplementing the results of prior studies (Romberg & Saffran, 2013; Vuong, Meyer, & Christiansen, 2016). Complementary findings are also observed in studies of children's nonword repetition using stimuli derived from natural language statistics: through exposure to a language, individuals pick up on lexical chunks (in this case, non-adjacent frames) and sub-lexical chunks (e.g., the bigram information within those frames), leading to enhanced recall of strings that follow these lexical and sub-lexical statistics (Jones, 2016). These findings also illustrate how the behaviors observed in the current paper tap into real-world psychological phenomena.

Contrary to the results of Endress and Mehler (2009), and in line with those of Perruchet and Poulin-Charronnat (2012), our 2AFC results show a significant difference between

endorsement of the words over phantom words on the word identification trials. Further, we extend these results by observing a comparable effect on the generalization trials. These trials in particular discount the idea that individuals encode rule-like representations regarding syllable position—participants show better endorsement of target over phantom items even though both follow the purported rule structure (e.g., A precedes C). Rather, individuals appear to encode specific A-C combinations, suggesting the acquisition of concrete items over abstract rules. These results thus lend important insights into the nature of exemplar-based learning and generalization, and how the two unfold over time: individuals represent learned items with enough flexibility to generalize over exemplars relatively early during the learning process.

## 4. General Discussion

Successful learning necessitates more than the encoding of specific items or events in the environment—it requires generalizing to novel instances. The current paper tested the question: can general-purpose statistical learning and memory processes account for the acquisition and generalization of non-local dependencies, a common challenge to many memory and exemplar-based learning models? Our results provide strong evidence for the statistically-based chunking of non-adjacent structure that is not reducible to positional encoding and does not require rule learning. The results also suggest that participants represent non-adjacent information as input-specific chunks that can scaffold structural generalization (Lieven, 2016).

In line with statistical learning-based theories (e.g., Aslin & Newport, 2012), our results suggest that structure learning and generalization may be more computationally unified in adults than previously assumed (Endress & Bonatti, 2016). Rather than these two abilities requiring distinct statistical and rule-like computations, they can instead rely on similar statistical learning

and memory mechanisms (e.g., Frost & Monaghan, 2016; Perruchet et al., 2004). Furthermore, while some theories posit that statistical information is insufficient for the acquisition of non-adjacent structures from speech, let alone generalization (Endress & Mehler, 2009; Endress, Scholl, & Mehler, 2005), we show that non-adjacent dependencies can be both acquired and generalized without the recruitment of additional cues, replicating the results of previous studies (Frost & Monaghan, 2016). Although the recruitment of additional cues can facilitate the acquisition of non-local dependency structures (e.g., de Diego Balauger, Rodriguez-Fornells, & Bachoud-Levi, 2015; de Diego Balauger, Toro, Rodriguez-Fornells, & Bachoud-Levi, 2007; Fló, 2021; Newport & Aslin, 2004; Rodriguez-Fornells, Cunillera, Mestres-Misse, & de Diego-Balauger, 2009; Van den Bos, Christiansen, & Misyak, 2012), they are not a strict requirement for learning (Frost et al., 2019; Onnis, Monaghan, Christiansen, & Chater, 2004).

Our results further elucidate how individuals move from exemplar-based learning to forming broader generalizations—a key area of inquiry in psychology. Rather than requiring the gradual formation of abstract rules, learners appear to chunk specific exemplars from the input based on adjacent and non-adjacent statistical regularities (e.g., *I__them*), which serve as a launchpad for generalizing beyond what was encountered (e.g., *I saw them, I like them, I want to eat them,* etc.). These results parallel usage-based frameworks of language acquisition (e.g., Tomasello, 2003; Lieven, 2016), which view productivity as arising from the interplay of encoding input-specific constructions and abstracting over them to create novel variations. For example, learners appear to extract and store lexical frames (Lieven, Behrens, Speares, & Tomasello, 2003; Lieven, Salomo, & Tomasello, 2003): specific multiword constructions that frequently occur in an input (e.g., *I want* __). Productivity occurs when learners insert a novel word or multiword chunk

into the empty slot of the frame (e.g., *I want this, I want to go home,* etc.), enabling learners to generalize over statistically-learned chunks.

A chunking-based computational model that discovers lexical frames may provide a window into how chunking could operate in the learning and generalization that took place in our experiments. This lexical frame discovery model involves a minor, principled extension to an earlier Chunk-Based Learner (CBL; McCauley & Christiansen, 2019a) model, which aimed to simulate children's language comprehension and production under real-time memory constraints. CBL was exposed to corpora of child-directed speech—one word at a time—using backwards transitional probabilities between words (which learners are sensitive to: e.g., Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008; Saffran, 2001; 2002) to decide whether to group words together as a chunk or not. In this way, the model processes the input incrementally, while building up an inventory of chunks that consist of one or more words. Through a simple generalization process, multiword chunks can then be used to facilitate further processing: previous encountered chunks are automatically grouped together independently of transitional probability information. As a model of early language acquisition, CBL was able to simulate the kind of shallow parsing that likely plays a role in children's language comprehension and the use of distributional regularities in their production of utterances. The model showed strong performance across a typologically diverse range of languages (McCauley & Christiansen, 2019a) while also capturing psycholinguistic data from both children (McCauley & Christiansen, 2014) and adults (Grimm, Cassani, Gillis, & Daelemans, 2017).

This lexical frame version of CBL (CBL+LF; McCauley & Christiansen, 2019b) incorporated a slight change to the generalization process: when the model has discovered five or more multiword chunks of the same size and which differ only by a single word (in the same

position), it creates a lexical frame with an empty slot. For example, if the model learns the chunks *on our own, on your own, on their own, on his own, on its own,* it automatically generalizes over them to create the lexical frame *on__own.*[2] These lexical frames are stored in the model's chunk inventory (or long-term memory), where they can combine with other words and chunks to produce novel utterances not encountered in the input (e.g., *on my own, on her own,* etc.). The model thus demonstrates how the combination of statistical learning and chunking mechanisms can capture the generalization of non-adjacent structures. These simulations, along with the data presented here, suggest that chunks do not only comprise contiguous units, but can also incorporate non-adjacent lexical frames wherein other chunks can be placed, thereby extending the results of prior memory models.

Importantly, these results show that chunk formation and statistical dependency learning are part and parcel of the same learning process. In Experiment 1, participants appear to form chunks of information based on the different statistics present in the input (Perruchet & Poulin-Charronnat, 2012; Slone & Johnson, 2015; 2018; Wang et al., 2019), encoding specific trigrams from the input as well as non-adjacent dependencies. Experiment 2 provided further evidence of this ability, by illustrating how participants encode non-adjacent chunks beyond positional information, though they still encode adjacent information as well (exemplified by the fact that participants recall the bigrams equally well in the target and phantom items). It may thus be the case that theories of statistical learning need not rule out the acquisition of transitional probabilities

[2] The CBL-LF model thus answers the call by Kol, Nir and Wintner (2014) for the kind of psychologically plausible, yet computationally rigorous, approximation of the Traceback method proposed by Lieven, Behrens, Speares, and Tomasello (2003) to capture children's item-based language learning. The threshold of 5+ for creating a lexical frame was chosen as a more conservative constraint than the 4+ used by Cameron-Faulkner, Lieven, and Tomasello (2003) in their hand-coded analysis of child-directed speech.

in favor of chunk information, or vice-versa—rather, learners appear to utilize both. Indeed, computational models that involve both statistical computation and chunk formation provide a stronger fit to statistical learning data than those that exclusively rely on transitional probability calculation (French et al., 2011; McCauley & Christiansen, 2019a, 2019b; Perruchet & Vinter, 1998), and is consistent with theories that view statistical learning as a suite of domain-general computations (Frost et al., 2015; 2019). While our data cannot determine whether chunk formation and statistical computation occur in parallel (McCauley & Christiansen, 2019a; 2019b), or if statistical sensitivity manifests due to chunking (Perruchet & Vinter, 1998; Perruchet & Pacton, 2006; Thiessen & Pavlik, 2013), they do suggest that both are required for learning and generalization.

Our results also contribute further insights to the memory literature, and particularly to studies employing serial recall. Individuals who have picked up on the statistical regularities of artificial languages show better recall of grammatical items, when controlling for baseline phonological working memory (Conway, Bauernschmidt, Huang & Pisoni, 2010; Isbilen et al., 2020; Kidd et al., 2020). Similarly, memory for sequences of high frequency words from natural language is superior to memory for strings of low frequency words (Hulme, Roodenrys, Schweickert, Brown, Martin, & Stuart, 1997), and long-term lexical and phonological knowledge facilitates recall, when test items are manipulated to leverage distributional regularities from an artificial language that participants were exposed to (Majerus, van Der Linden, Mulder, & Peters, 2004). Other studies have shown that when individuals are trained to associate pairs of words, these pairs are later treated as a single chunked unit (Cowan, Chen, & Rouder, 2004), and classic memory studies show that word predictability and frequency facilitate recall (Baddeley, Conrad, & Hull, 1965). While prior observations have typically been limited to facilitation from adjacent

statistical information, we extend these findings here by showing comparable boosts to memory performance from the statistical learning of non-adjacent information.

The question of how non-adjacent information may be stored in memory is a topic of some speculation. For example, in the visual domain, some studies suggest that individuals utilize such structural dependencies to simplify their mental representation of stimuli to allow their retention in memory. This is thought to take different forms, with individuals encoding both detailed information about specific items and the overall summary of a scene (Brady & Tenenbaum, 2013; Hollingworth & Henderson, 2003; Oliva, 2005), or by regularizing the features of a scene to facilitate compressibility and reduce short-term memory load—though such compression may also cause memory errors by oversimplifying the data (Lazartigues, Lavigne, Aguilar, Cowan, & Mathy, 2021). Yet other studies have challenged the notion of chunking as a form of data compression that frees up working memory (Norris, Kalm, & Hall, 2020), and instead suggest that chunking may be achieved by redintegration. In this account, chunked representations only reside in long-term memory, enabling individuals to rebuild whole representations from degraded traces in short-term memory (but see Brady, Konkle, & Alvarez, 2009; Thalmann, Souza, & Oberauer, 2019, for evidence of compression in memory). While the current paper was not designed to disentangle the compression and redintegration accounts, our data do suggest that non-adjacent statistical information appears to facilitate memory in a similar fashion as adjacent information: by allowing the cognitive system to build larger units of representation, and thereby decreasing the number of items that need to be held in working memory. This is consistent with chunking models of serial recall (e.g., Cowan et al., 2012), which show that working memory limitations interact with long term memory. While this and other models have primarily focused on items

linked by adjacent probabilities, their results may extend to items comprising non-adjacent regularities as well.

Additionally, acquisition and generalization appear to occur in parallel, rather than requiring mastery of the language before evidence of generalization can be observed, as has been previously suggested. However, while these two abilities appear to emerge around the same time at test, we acknowledge the limitations of the current study in providing online data that tracks the time course of learning during training, or the specific mechanisms involved therein. Just as decades of 2AFC results are taken to be indicative of the calculation of transitional probabilities during statistical learning, we predict that the evidence of chunk formation on SICR may similarly indicate the involvement of chunking during learning, and the apparent concurrence of learning and generalization.

Furthermore, our data do not speak to how word learning and generalization proceed in infants, nor its developmental trajectory (Gómez & Maye, 2005). Evidence for how these abilities unfold in infants is currently mixed. Indeed, it has previously been suggested that generalization may only appear later in development—and only then with the incorporation of additional acoustic cues (Marchetto & Bonatti, 2013). Yet the opposite has been reported for the acquisition of musical sequences (Dawson & Gerken, 2009), with four-month-old infants successfully generalizing tone and chord combinations that follow a regular pattern, but not seven-month-olds. By contrast, Frost et al. (2020) show that seventeen-month-olds both segment and generalize structure after a brief period of exposure to an artificial speech stream on the basis of statistical cues alone. Similar results have also been observed for visual statistical learning (Saffran, Pollak, Seibel, & Shkolnik, 2007). While these results make it difficult to untangle the precise developmental timeline of generalization, collectively, they do suggest that it occurs for both linguistic and non-linguistic

stimuli. As this ability applies across domains, it thus may be underpinned by general cognitive rather than language-specific mechanisms.

The processes involved in learning and generalization have long been debated. Here, we suggest that the process characterized as statistical learning may involve a whole host of domain-general computations working in parallel (Frost et al., 2015; 2019), among which chunking plays a central role. This view differs somewhat from other accounts that propose multiple task-specific mechanisms (the "more than one mechanism", or MOM hypothesis; Endress & Bonatti, 2007; 2016), where statistical computations calculate transitional probabilities among adjacent and non-adjacent information in speech, while a secondary mechanism extracts rules. Instead, we suggest that statistically-facilitated chunking works in tandem with other domain-general processes to enable the learning and generalization of structure in memory, language, and other aspects of cognition.

## 5. Context of the research

While chunking models are powerful in capturing the learning of temporally and spatially contiguous information across domains, how such memory mechanisms might apply to remote dependencies has remained a relative mystery. Research into this area provides fertile ground for investigating the nature of exemplar-based learning and how individuals form generalizations over items—the representational foundations of vocabulary and grammar in the psychology of language and in cognitive science at large. Prior evidence shows that SICR is a highly sensitive tool for investigating learning and representation of adjacent dependencies. The current study was

motivated by the question of whether SICR could be expanded to non-local dependencies, which would allow for a broader notion of chunking. Our results thus unlock a trove of possibilities for studying learning and generalization in a range of participants and tasks. Fruitful future directions may include investigating how representations change with age (e.g., are children more flexible in their acquisition or generalization of structures than adults?), how children with language disorders perform on tasks involving the chunking of adjacent and non-adjacent structures, whether skill in one statistical learning task correlates with skill in another, whether the learning of artificial dependencies correlates with the kinds of information individuals can learn in the real world, and how statistical learning relates to other aspects of cognition, such as auditory and visual perception, and memory and attention systems. Such avenues may in turn enable us to bridge statistically-based chunking in language to a host of broad phenomena in cognition.

## References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science, 21*, 170-176.

Baddeley, A. D., Conrad, R., & Hull, A. J. (1965). Predictability and immediate memory for consonant sequences. *Quarterly Journal of Experimental Psychology*, *17*, 175-177.

Baddeley, A., Papagno, C., & Andrade, J. (1993). The sandwich effect: The role of attentional factors in serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 862-870.

Bauer, P. J., & Dow, G. A. (1994). Episodic memory in 16-and 20-month-old children: Specifics are generalized but not forgotten. *Developmental Psychology*, *30*, 403-417.

Black, A. W., Taylor, P., & Caley, R. (1990). *The festival speech synthesis system*. Edinburgh, UK: Centre for Speech Technology Research (CSTR), University of Edinburgh.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2006). How to hit Scylla without avoiding Charybdis: comment on Perruchet, Tyler, Galland, and Peereman (2004). *Journal of Experimental Psychology: General, 135*, 314-321.

Botvinick, M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 351-358.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27*, 843-873. https://doi.org/10.1207/s15516709cog2706_2

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Brady, T.F., & Tenenbaum, J.B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review, 120,* 85–109.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*, 487.

Chen, Z., & Cowan, N. (2009). Core verbal working memory capacity: The limit in words retained without covert articulation. *Quarterly Journal of Experimental Psychology, 62*, 1420-1429.

Christiansen, M. H. (2019). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science, 11,* 468-481. https://doi.org/10.1111/tops.12332.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39,* e62.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences, 3*, 1-15.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114*, 356-371.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-114.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science, 15*, 634-640.

Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review, 119*, 480-499.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition, 111,* 378-382.

De Diego-Balaguer, R., Rodriguez-Fornells, A. & Bachoud-Lévi, A. C. (2015). Prosodic cues enhance rule learning by changing speech segmentation mechanisms, *Frontiers in Psychology, 6,* 1478.

De Diego-Balauger, R., Toro, J. M., Rodriguez-Fornells, A., & Bachoud-Levi, A. C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS One, 2,* p. 01175.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*, 1136-1146.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*, 247-299.

Endress, A. D., & Bonatti, L. L. (2016). Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*, 19-35.

Endress, A. D., Cahill, D., Block, S., Watumull, J., & Hauser, M. D. (2009). Evidence of an evolutionary precursor to human language affixation in a non-human primate. *Biology Letters, 5,* 749-751.

Endress, A.D. & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351-367.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348-353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Psychology: General, 134*, 406-419.

Fló, A. (2021). Evidence of ordinal position encoding of sequences extracted from continuous speech. *Cognition*, 104646.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107-125.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614-636.

Frost, R., Armstrong, B.C. & Christiansen, M.H. (2019). Statistical learning research: A critical review and possible directions. *Psychological Bulletin, 145*, 1128-1153.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, *19*, 117-125.

Frost, R.L.A., Isbilen, E.S., Monaghan, P. & Christiansen, M.H. (2019). Testing the limits of non-adjacent dependency learning: statistical segmentation and generalization across domains. In A. Goel, C. Seifert & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1787-1793). Austin, TX: Cognitive Science Society.

Frost, R. L. A., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., Rowland, C. F., & Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology, 120*, 101291.

Frost, R.L.A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech, *Cognition*, *147*, 70- 74.

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

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*, 260-272.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, *6*, 225-255.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-436.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109-135.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*, 183-206.

Goodsitt, J., Morse, P., Ver Hoeve, J., & Cowan, N. (1984). Infant speech recognition in multisyllabic contexts. *Child Development*, *55*, 903-910.

Greene, R. L., Elliot, C. L., & Smith, M. D. (1988). When do in- terleaved suffixes improve recall? *Journal of Memory and Lan- guage*, *27*, 560-571.

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology*, *8*, 555. https://doi.org/10.3389/fpsyg.2017.00555

Hitch, G. J. (1975). The role of attention in visual and auditory suffix effects. *Memory and Cognition, 3,* 501-505.

Hockett, C. F. (1959). Animal "languages" and human language. *Human Biology, 31,* 32-39.

Hollingworth, A., & Henderson, J. M. (2003). Testing a conceptual locus for the inconsistent object change detection advantage in real-world scenes. *Memory & Cognition, 31,* 930–940.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., 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,* 1217-1232.

Isbilen, E. S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2021). Online repository: Statistically-based chunking of non-adjacent dependencies. Open Science Framework. URL: https://osf.io/k8mys/.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically-induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science, 44,* e12848.

Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. *Frontiers in Psychology, 3,* 167. doi: 10.3389/fpsyg.2012.00167.

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition, 53,* 79–88.

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: the case of nonword repetition. *Developmental Science, 17*, 298-310.

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition, 144*, 1-13.

Kidd, E., Arciuli, J., Christiansen M.H., Isbilen E.S., Revius, K. & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology, 200*, 104964.

Kinder, A., & Assmann, A. (2000). Learning artificial grammars: No evidence for the acquisition of rules. *Memory & Cognition, 28*, 1321-1332.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 79-91.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 169-181.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language, 41*, 176-199. https://doi.org/10.1017/S0305000912000694

Kuhn, G., & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: implicitly learning more than chunks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1417-1432.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software, 82*, 1–26. doi: 10.18637/jss.v082.i13.

Lazartigues, L., Lavigne, F., Aguilar, C., Cowan, N., & Mathy, F. (2021). Benefits and pitfalls of data compression in visual working memory. *Attention, Perception, and Psychophysics, 83*, 2843–2864. https://doi.org/10.3758/s13414-021-02333-x

Lieven, E. (2016). Usage-based approaches to language development: Where do we go from here? *Language and Cognition*, *8*, 346-368.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333-370.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187-219.

Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: a usage-based analysis. *Cognitive Linguistics, 20*, 481–507.

Majerus, S., van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language, 51*, 297-306.

Marchetto, E., & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, *67*, 130-150.

Marchetto, E. & Bonatti, L. L. (2015). Finding words and word structure in artificial speech: the development of infants' sensitivity to morphosyntactic regularities. *Journal of Child Language, 42*, 873-902. doi:10.1017/S0305000914000452

McCauley, S.M. & Christiansen, M.H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9*, 419-436. https://doi.org/10.1075/ml.9.3.03mcc

McCauley, S. M., & Christiansen, M. H. (2019a). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126,* 1-51. https://doi.org/10.1037/rev0000126

McCauley, S.M. & Christiansen, M.H. (2019b). Modeling children's early linguistic productivity through the automatic discovery and use of lexically-based frames. In A. Goel, C. Seifert & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 782-788). Austin, TX: Cognitive Science Society.

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

Murray, D. J. (1966). Intralist interference and rehearsal time in short-term memory. *Canadian Journal of Psychology, 20,* 413- 426.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance. I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48,* 127–162.

Nicholls, A. P., Parmentier, F. B., Jones, D. M., & Tremblay, S. (2005). Visual distraction and visuo-spatial memory: A sandwich effect. *Memory, 13,* 357-363.

Norris, D.G., & Kalm, K., & Hall, J. (2020). Chunking and redintegration in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 872–893.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 251–256). San Diego, CA: Elsevier.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in non-adjacent

dependencies. *Proceedings of the 26th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.*

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*, 225-237.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745-2750.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80,* 674-685.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*, 604-607.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in Cognitive Science, 11,* 520-535.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition, 36,* 1299-1305.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*, 233–238.

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*, 807-818.

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: no need for algebraic-like computations. *Journal of Experimental Psychology: General*, *133*, 573-583.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39,* 246-263.

Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature, 431,* 768-774.

Radulescu, S., Wijnen, F., & Avrutin, S. (2020). Patterns bit by bit: An entropy model for rule induction. *Language Learning and Development, 16,* 109-140.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rodriguez-Fornells, A., Cunillera, T., Mestres-Misse, A., & de Diego-Balauger, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*(1536), 3711-3734.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science, 37,* 1290-1320.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81,* 149-169.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47,* 172-196.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1928.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition, 105,* 669-680.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools, Inc.

Siegelman, N., Bogaerts, L., Armstrong, B. C., & Frost, R. (2019). What exactly is learned in visual statistical learning? Insights from Bayesian modeling. *Cognition, 192,* 104002.

Slone, L., & Johnson, S. P. (2015). Statistical and chunking processes in adults' visual sequence learning. In D. C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2218–2223). Austin, TX: Cognitive Science Society.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition, 178,* 92-102.

Thalmann, M., Souza, A., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45,* 37–55.

Thiessen, E. D., & Pavlik, P. I. (2013). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science, 37,* 310-343.

Tomasello, M. (2003). Introduction: Some surprises for psychologists. In M. Tomasello (Ed.), *New psychology of language: Cognitive and functional approaches to language structure* (pp. 1–14). Mahwah, NJ: Lawrence Erlbaum.

Trotter, A. S., Monaghan, P., Beckers, G. J., & Christiansen, M. H. (2020). Exploring Variation Between Artificial Grammar Learning Experiments: Outlining a Meta-Analysis Approach. *Topics in Cognitive Science, 12,* 875–893.

Van den Bos, E., Christiansen, M.H. & Misyak, J.B. (2012). Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *Journal of Memory and Language, 67,* 507-520.

Vuong, L. C., Meyer, A. S., & Christiansen, M. H. (2016). Concurrent statistical learning of adjacent and nonadjacent dependencies. *Language Learning, 66,* 8-30.

Wang, F. H., Zevin, J., & Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cognitive Psychology, 113*, 101223. https://doi.org/10.1016/j.cogpsych.2019.101223

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *43*, 865-878.

Wilson, B., Spierings, M., Ravignani, A., Mueller, J.L., Mintz, T.H., Wijnen, F., Van der Kant, A., Smith, K. and Rey, A., 2020. Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science*, *12*, 843-858.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication*, *2*, 57-89.

# Appendix

*Appendix 1: Input and generalization words for Experiments 1& 2*

Appendix 1a

*Table of input words*

| Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|
| dufoki | befoki | dufoga | dufoki |
| duliki | beliki | duliga | duliki |
| duraki | beraki | duraga | duraki |
| gafobe | dufopu | kifobe | pufobe |
| galibe | dulipu | kilibe | pulibe |
| garabe | durapu | kirabe | purabe |
| tafopu | tafoga | pufota | tafoga |
| talipu | taliga | pulita | taliga |
| tarapu | taraga | purata | taraga |

Appendix 1b

*Table of generalization words*

| Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|
| duthiki | bethiki | duthiga | duthiki |
| duveki | beveki | duvega | duveki |
| duzoki | bezoki | duzoga | duzoki |
| gathibe | duthipu | kithibe | puthibe |
| gavebe | duvepu | kivebe | puvebe |
| gazobe | duzopu | kizobe | puzobe |
| tathipu | tathiga | puthita | tathiga |
| tavepu | tavega | puveta | tavega |
| tazopu | tazoga | puzota | tazoga |

*Appendix 2: Test items for Experiment 1*

Appendix 2a

*SICR items for Experiment 1*

| Input word or Generalization item | Target or foil | Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|---|---|
| w | t | dulikitafopu | duliputafoga | pulitakifobe | pulibedufoki |
| w | t | durakigafobe | durapubefoki | puratadufoga | purabetafoga |
| w | t | talipugarabe | taligaberaki | kilibeduraga | dulikitaraga |
| w | t | tarapugalibe | taragabeliki | kirabeduliga | durakitaliga |
| w | t | tafopuduliki | tafogadulipu | kifobepulita | dufokipulibe |
| w | t | galibeduraki | belikidurapu | duligapurata | taligapurabe |
| w | t | garabedufoki | berakidufopu | duragapufota | taragapufobe |
| w | t | gafobetalipu | befokitaliga | dufogakilibe | tafogaduliki |
| w | f | gabepulirata | bekigalirata | dugabeliraki | tagakiliradu |
| w | f | litapukidofu | litagapufodu | likibetafopu | lidukibefopu |
| w | f | foraduputaki | foradugatapu | forapubekita | forapukidube |
| w | f | kiberaligadu | pukiralibedu | tagaralidupu | begaralitapu |
| w | f | pubelifogata | gakilifobeta | begalifoduki | kigalifotadu |
| g | t | duvekitathipu | duveputathiga | puvetakithibe | puvebeduthiki |
| g | t | duzokigavebe | duzopubeveki | puzotaduvega | puzobetavega |
| g | t | duthikigazobe | duthipubezoki | puthitaduzoga | puthibetazoga |
| g | t | tavepugathibe | tavegabethiki | kivebeduthiga | duvekitathiga |
| g | t | tazopuduthiki | tazogaduthipu | kizobeputhita | duzokiputhibe |
| g | t | tathipuduveki | tathigaduvepu | kithibepuveta | duthikipuvebe |
| g | t | gavebetazopu | bevekitazoga | duvegakizobe | tavegaduzoki |
| g | t | gazobetavepu | bezokitavega | duzogakivebe | tazogaduveki |
| g | t | gathibeduzoki | bethikiduzopu | duthigapuzota | tathigapuzobe |
| g | f | kipuvethitadu | pugavethitadu | tabevethikipu | bekivethidupu |
| g | f | bekithizoduga | kiputhizodube | gatathizopudu | gabethizoputa |
| g | f | dubezovegaki | dukizovebepu | pugazoveduta | pugazovetabe |
| g | f | vetabegapuzo | vetakibegazo | vekigadubezo | vedugatakizo |

Appendix 2b

*2AFC foil items for Experiment 1*

| Input word or Generalization foil | Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|---|
| w | bedufo | fokibe | bepura | bedura |
| w | fobega | gadura | fogadu | betafo |
| w | kigafo | kidufo | gapufo | fogata |
| w | kitara | ligabe | libedu | gapufo |
| w | libedu | likidu | ligapu | kipura |
| w | likita | liputa | litaki | libedu |
| w | lipuga | pubefo | rabepu | ligapu |
| w | pudura | putara | tadufo | likita |
| w | rapudu | ragadu | takira | rakipu |
| g | begave | fothibe | bekizo | beputhi |
| g | fothiga | gatazo | fothidu | fothita |
| g | kiduthi | kibeve | gaduve | gatave |
| g | liveta | liveta | liveki | kiduzo |
| g | putazo | puduthi | razopu | livedu |
| g | razodu | razodu | taputhi | razopu |
| g | thiputa | thigata | thibeki | thikidu |
| g | vekidu | vepudu | vetapu | vebepu |
| g | zobega | zokibe | zogadu | zogata |

*Appendix 3: Test items for Experiment 2*

Appendix 3a

*SICR items for Experiment 1*

| Input word or Generalization item | Target or foil | Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|---|---|
| w | t | talipugarabe | taligaberaki | kilibeduraga | dulikitaraga |
| w | t | gafobetalipu | befokitaliga | dufogakilibe | tafogaduliki |
| w | t | tafopuduliki | tafogadulipu | kifobepulita | dufokipulibe |
| w | t | garabedufoki | berakidufopu | duragapufota | taragapufobe |
| w | t | tarapugalibe | taragabeliki | kirabeduliga | durakitaliga |
| w | t | galibeduraki | belikidurapu | duligapurata | taligapurabe |
| w | t | dulikitafopu | duliputafoga | pulitakifobe | pulibedufoki |
| w | t | durakigafobe | durapubefoki | puratadufoga | purabetafoga |
| w | f | talibegarapu | talikiberaga | kiligadurabe | duligataraki |
| w | f | galikitafobe | beliputafoki | dulitakifoga | talibedufoga |
| w | f | dulipugafoki | duligabefopu | pulibedufota | pulikitafobe |
| w | f | dufoputaraki | dufogatarapu | pufobekirata | pufokidurabe |
| g | t | duzokigavebe | duzopubeveki | puzotaduvega | puzobetavega |
| g | t | gafibeduzoki | befikiduzopu | dufigapuzota | tafigapuzobe |
| g | t | tazopudufiki | tazogadufipu | kizobepufita | duzokipufibe |
| g | t | gazobetavepu | bezokitavega | duzogakivebe | tazogaduveki |
| g | t | duvekitafipu | duveputafiga | puvetakifibe | puvebedufiki |
| g | t | gavebetazopu | bevekitazoga | duvegakizobe | tavegaduzoki |
| g | t | tafipuduveki | tafigaduvepu | kifibepuveta | dufikipuvebe |
| g | t | dufikigazobe | dufipubezoki | pufitaduzoga | pufibetazoga |
| g | f | tazokiduvepu | tazopuduvega | kizotapuvebe | duzobepuveki |
| g | f | gazopudufibe | bezogadufiki | duzobepufiga | tazokipufiga |
| g | f | tavebegafipu | tavekibefiga | kivegadufibe | duvegatafiki |
| g | f | duzobetafiki | duzokitafipu | puzogakifita | puzogadufibe |

Appendix 3b

*2AFC phantom word foils items for Experiment 2*

| Input word or Generalization foil | Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|---|
| w | dulipu | duliga | pulibe | puliki |
| w | durabe | duraki | puraga | puraga |
| w | dufopu | dufoga | pufobe | pufoki |
| w | talibe | taliki | kiliga | duliga |
| w | taraki | tarapu | kirata | durabe |
| w | tafobe | tafoki | kifoga | dufoga |
| w | galiki | belipu | dulita | talibe |
| w | garapu | beraga | durabe | taraki |
| w | gafoki | befopu | dufota | tafobe |
| g | duvepu | duvega | puvebe | puveki |
| g | duzobe | duzoki | puzoga | puzoga |
| g | duthibe | duthiki | puthiga | puthiga |
| g | tavebe | taveki | kivega | duvega |
| g | tazoki | tazopu | kizota | duzobe |
| g | tathiki | tathipu | kithita | duthibe |
| g | gaveki | bevepu | duveta | tavebe |
| g | gazopu | bezoga | duzobe | tazoki |
| g | gathipu | bethiga | duthibe | tathiki |

## Figures

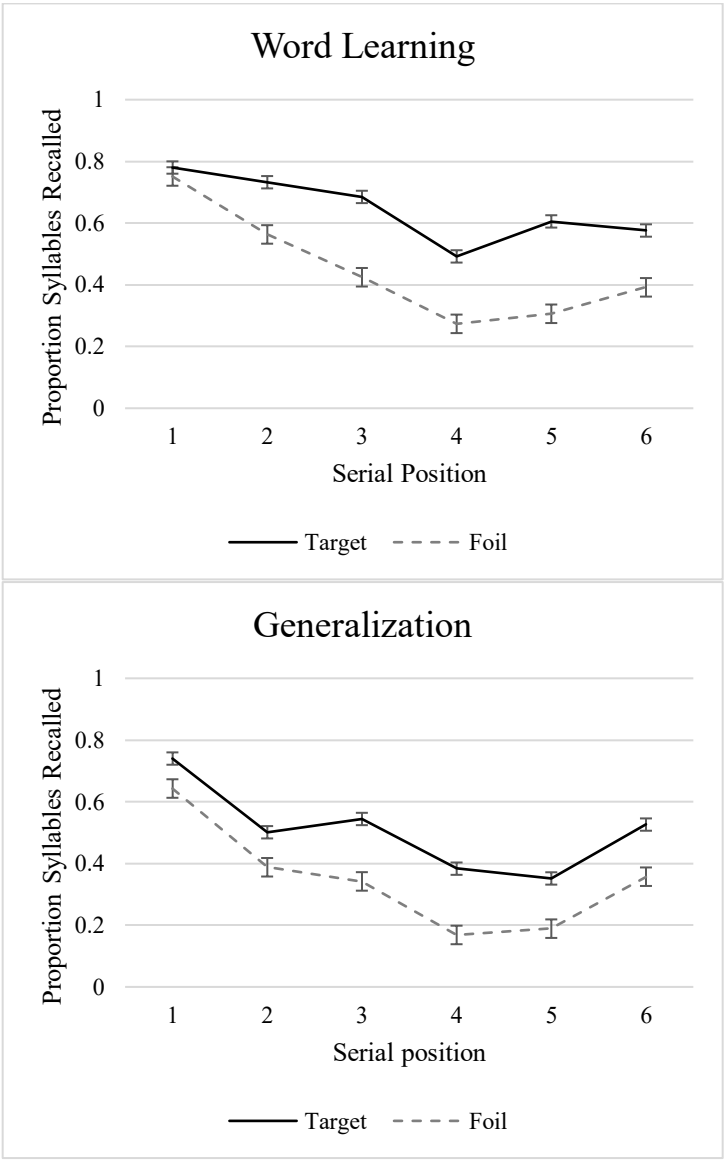


Figure 1. Serial position curves showing the accuracy of recall on the SICR task. Accuracy was higher for the target items, which follow the statistics of the artificial language, than for the foil items, both on the word learning and generalization trials. Error bars reflect standard error.

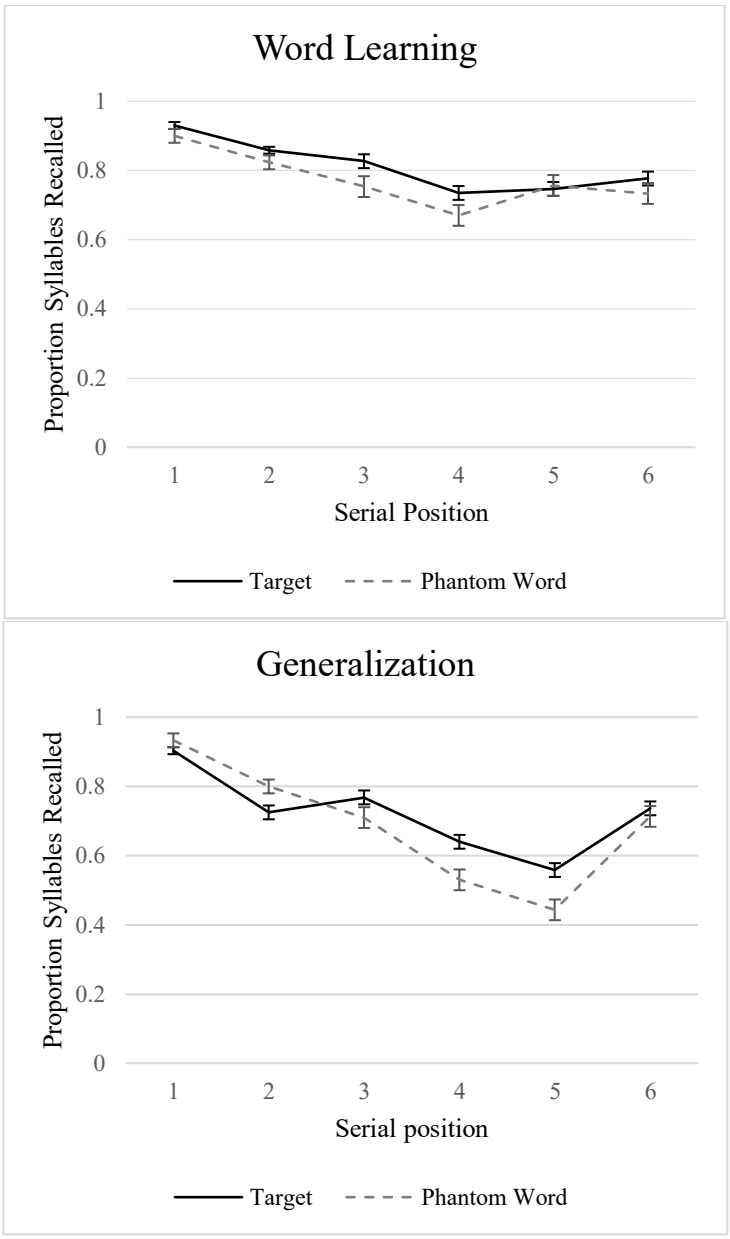


Figure 2. SICR serial position curves for the word learning and generalization trials by item type (target vs. phantom word strings). Error bars reflect standard error.