Cognition 225 (2022) 105123

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short communication

Individual differences in artificial and natural language statistical learning

Erin S. Isbilen [a,b,*], Stewart M. McCauley [c], Morten H. Christiansen [a,b,d]

[a] Cornell University, Department of Psychology, USA
[b] Haskins Laboratories, USA
[c] University of Iowa, Department of Communication Sciences and Disorders, USA
[d] Aarhus University, Interacting Minds Centre and School of Communication and Culture, Denmark

ARTICLE INFO

Keywords:
Chunking
Statistical learning
Individual differences
Memory
Language acquisition

ABSTRACT

Statistical learning (SL) is considered a cornerstone of cognition. While decades of research have unveiled the remarkable breadth of structures that participants can learn from statistical patterns in experimental contexts, how this ability interfaces with real-world cognitive phenomena remains inconclusive. These mixed results may arise from the fact that SL is often treated as a general ability that operates uniformly across all domains, typically assuming that sensitivity to one kind of regularity implies equal sensitivity to others. In a preregistered study, we sought to clarify the link between SL and language by aligning the type of structure being processed in each task. We focused on the learning of trigram patterns using artificial and natural language statistics, to evaluate whether SL predicts sensitivity to comparable structures in natural speech. Adults were trained and tested on an artificial language incorporating statistically-defined syllable trigrams. We then evaluated their sensitivity to similar statistical structures in natural language using a multiword chunking task, which examines serial recall of high-frequency word trigrams—one of the building blocks of language. Participants' aptitude in learning artificial syllable trigrams positively correlated with their sensitivity to high-frequency word trigrams in natural language, suggesting that similar computations span learning across both tasks. Short-term SL taps into key aspects of long-term language acquisition when the statistical structures—and the computations used to process them—are comparable. Better aligning the specific statistical patterning across tasks may therefore provide an important steppingstone toward elucidating the relationship between SL and cognition at large.

1. Introduction

Statistical learning (SL) is a foundational building block of human cognition. Recent years have seen an upswell of studies evaluating the connection between the learning of statistical patterns and a multitude of high-level cognitive abilities (and disabilities) across domains (Bogaerts, Siegelman, Christiansen, & Frost, 2022), particularly in language. This interest in individual differences is in part predicated on one of the central premises of SL experiments: that they simulate how learning unfolds in the real world. These studies are treated as a window onto how linguistic skills progress more broadly, and critically inform theories of language development, usage, and evolution.

Currently, however, direct evidence bridging SL and language remains inconclusive. Although some studies have unearthed positive correlations between SL and language aptitude, others report little to no correlation. For example, several experiments find that SL predicts reading (Arciuli & Simpson, 2012) and sentence processing (Misyak & Christiansen, 2012), while others display null correlations with related skills (e.g., Haebig, Saffran, & Ellis Weismer, 2017; Siegelman & Frost, 2015; van Witteloostuijn, Boersma, Wijnen, & Rispens, 2021). These conflicting results pose a fundamental challenge to our understanding of SL's role in language, and in cognition in general: if SL studies do in fact tap into mechanisms involved in real-world learning, then reliable relationships between the two would be expected (Kidd, Donnelly, & Christiansen, 2018).

The mixed correlations between SL and other cognitive functions derive from a variety of sources. Perhaps the most formative of these is how the psychological sciences have historically conceptualized SL: as a single, general-purpose mechanism that is deployed across all cognitive arenas independent of the statistics involved. However, sensitivity to one kind of statistical information does not necessitate equal proficiency in processing another. Striking differences in SL across modalities and domains have been documented (e.g., Conway & Christiansen, 2005; Siegelman & Frost, 2015), as well as across statistical dependencies

* Corresponding author at: Haskins Laboratories 300 George Street, #900, New Haven, CT 06511, USA.
E-mail address: erin.isbilen@yale.edu (E.S. Isbilen).

https://doi.org/10.1016/j.cognition.2022.105123
Received 4 June 2021; Received in revised form 18 February 2022; Accepted 5 April 2022
0010-0277/Published by Elsevier B.V.

within domains (e.g., Trotter, Monaghan, Beckers, & Christiansen, 2020). Indeed, even within individuals, sensitivity to adjacent regularities does not predict sensitivity to non-adjacent regularities in the same modality and domain (Siegelman & Frost, 2015), suggesting some degree of structure-related specificity of statistical computations even among patterns that are closely related. Furthermore, the regularities most relevant to navigating the natural environment remain poorly understood, making it difficult to deduce what constitutes a "good" statistical learner—whether shared statistical computations are deployed across distinct cognitive domains, or if they vary according to context (see Bogaerts et al., 2022, for a review).

Though often implicit, the theoretical assumption of SL as a single all-purpose mechanism has far-reaching repercussions for how the link between SL and learning in the real world is typically probed. While most SL studies present a single type of dependency such as adjacent or non-adjacent regularities, the natural language tasks that SL is compared against typically involve sensitivity to a much broader assortment of regularities. For example, many studies have focused on SL's connection with broad language-related cognitive abilities such as literacy (Qi, Sanchez Araujo, Georgan, Gabrieli, & Arciuli, 2019) or syntax (Kidd & Arciuli, 2016), despite mounting evidence suggesting the structure-dependent specificity of SL computations. A critical first step to establishing more robust theoretical links between SL and learning "in the wild" may therefore lie in determining whether sensitivity to specific kinds of artificial statistics predicts sensitivity to analogous statistics in natural language.

In the current paper, we tested whether sensitivity to syllable trigrams in the classic Saffran, Aslin, and Newport (1996) SL paradigm predicts sensitivity to high-frequency word trigrams in natural language. We chose combinations of three consecutive words as our natural language targets because such multiword chunks have been proposed as key building blocks of language learning and use (Goldberg, 2006; Lieven, Pine, & Baldwin, 1997), in both first and second language acquisition (Arnon & Christiansen, 2017; Ellis, 2012). Multiword chunks consisting of two or more words can be derived from the statistical properties of language (e.g., McCauley & Christiansen, 2019a), enabling the discovery of phrases and phrase fragments (e.g., *have to eat*) and the ability to generalize across them (e.g., *have to* __ $\geq$ *have to go, have to leave,* etc.; McCauley & Christiansen, 2019b). In this way, multiword sequences lay the foundation for many higher-level language skills including comprehension and production (Arnon & Snider, 2010; Bannard & Matthews, 2008; McCauley et al., 2021). Multiword chunks thus point to a kind of statistical structure that is highly relevant to natural language, and which may draw on similar statistical computations as those leveraged in SL experiments.

In the current experiment, we investigated the connection between artificial and natural language SL to determine whether different tests of in-lab SL vary in their ability to predict long-term distributional sensitivity. We exposed participants to an artificial language, then tested learning using the classic 2AFC task and the statistically-induced chunking recall task (SICR; Isbilen, McCauley, Kidd, & Christiansen, 2020), which gauges SL by comparing participants' serial recall of syllable strings that either adhere to or violate the statistics of the artificial language.[1] We then employed a multiword chunking task (MWC; McCauley, Isbilen, & Christiansen, 2017) to measure individuals' sensitivity to similar statistical patterns in natural language, accrued over many years of linguistic experience. In MWC, participants recall 12-word-long strings—a formidable challenge to typical working memory limitations (4 $\pm$ 1 items; Cowan, 2001). These strings are either composed of four high-frequency word trigrams (three-word combinations from natural language), or the same words presented in a random order. MWC can be construed as a measure of natural language SL: participants should perform better on the statistically-derived items if they have successfully acquired these trigram word co-occurrences from natural language, in line with growing literature on the importance of such multiword units in language acquisition and processing (see Christiansen & Arnon, 2017, for a review).[2]

We hypothesized that individual differences in sensitivity to artificial syllable trigrams would predict sensitivity to comparable trigram word structures from natural language. However, we expected that only SICR would correlate with natural language SL, given recent work demonstrating its superior reliability in measuring SL relative to 2AFC in both adults (Isbilen et al., 2020) and children (Kidd et al., 2020).[3] If confirmed, these results would provide key evidence to establishing the connection between SL and natural language, and lay the foundation for future studies clarifying the contribution of SL to broader cognitive functions.

2. Method

2.1. Participants

As was preregistered, 70 participants were recruited from Prolific (prolific.co). This number was based on a power analysis with an estimated effect size of $d = 0.4$, power $= 0.9$, and $p = .05$. Five participants were excluded due to technical errors/failure to complete the experiment. The final analyses were conducted on the remaining 65 participants (39 females/22 males/4 nonbinary; M age $= 20.81$, range $=$ 18–30, $SD = 2.08$). All were native speakers of American English, and were compensated with monetary payment.[4]

2.2. Materials

To measure artificial language SL,[5] the same language from Isbilen et al. (2020) was used, which consisted of six syllable trigrams/words (*tagalu, lomari, topoka, latibi, modipa, kibudu*). In addition, six 2AFC foils were created by randomizing the syllables of the target words, avoiding transitional probabilities from the language (*dikabi, lopadu, polubu, kigala, mamoti, tatori*). For SICR, 36 items (18 target/18 random) were created. The target trials comprised two-word combinations from the language (e.g., *tagalulomari*), and the random trials presented the same syllables in a randomized order that avoided reusing transitional probabilities from the language and 2AFC foils (e.g., *rilobimatila*).

The 20 MWC items (10 target/10 random) were adapted from McCauley et al. (2017) and Jolsvai, McCauley, and Christiansen (2020).

[1] The use of serial recall to measure statistical sensitivity is motivated by the fact that performance on such memory tasks is fundamentally shaped by distributional learning. For example, classic studies show that participants exhibit enhanced recall of sequences composed of high-frequency English word transitions compared to sequences comprising low-frequency transitions (Miller & Selfridge, 1951). Similarly, consonant strings containing high-frequency letter transitions are recalled better than those containing low-frequency transitions (Baddeley et al., 1965), with comparable memory facilitation observed from high-frequency digit sequences (Jones & Macken, 2015), and from artificial grammar statistics (Conway, Bauernschmidt, Huang, & Pisoni, 2010).

[2] MWC, as with any other natural language measure, likely captures more than statistical sensitivity alone, including aspects of semantics (Jolsvai et al., 2020) and syntax. However, given that previous SL studies have revealed null correlations with other measures of semantics and syntax, we hypothesized that the alignment of specific computational structures between the artificial and natural language measures might provide clearer correlations.

[3] While the fact that both MWC and SICR are recall tasks may play a role in any observed correlation, previous work on MWC demonstrates that it does not necessarily correlate with other recall tasks such as nonword repetition (McCauley et al., 2017). See Section 3 for further discussion.

[4] As per the Prolific study platform guidelines at the time of testing, participants were compensated $9.60/h.

[5] An extended methods section detailing stimulus creation, presentation, and data processing prior to analysis is reported in the Supplemental Materials.

Target items consisted of high-frequency word trigrams, extracted from a combination of the Fisher (Cieri, Graff, Kimball, Miller, & Walker, 2005) and American National corpora (Reppen, Ide, & Suderman, 2005), comprising 39 million words of American English. Each trigram was non-idiomatic and possessed an average frequency of 0.73/million words. Four high-frequency word trigrams were concatenated to form each target item (e.g., *have to eat good to know don't like them is really nice*). The same 12 words from each string were randomized to create the foils, avoiding high-frequency bigrams and trigrams (e.g., *really them nice have eat know to don't good like is to*).

All stimuli were synthesized using Google Text-to-Speech.[6] Transcriptions of the SICR and MWC items are reported in the Appendix.[7]

2.3. Procedure

To create the artificial language, 96 blocks, each containing a randomization of the six words mentioned above, were concatenated into an 11-min-long file. Participants were instructed to listen to the language carefully, and pay attention to the structures it may contain.

Following exposure, 2AFC was administered. On each trial, participants heard a target word and foil, and selected the one they remembered from training. Each target word was presented once with each foil, yielding 36 trials.

For SICR, the 36 strings described above were presented for serial recall. On each trial, participants heard a six-syllable-long string, after which a text box appeared, prompting them to type their response. Participants were not informed of the items' statistical structure: they were simply asked to remember each syllable to the best of their ability and type them in the correct order (Isbilen et al., 2020). To reduce spelling-related variability in their responses, participants were presented with a chart depicting the transcription of the syllables in the language before the task.

For MWC, a similar method to SICR was adopted. Participants were instructed to listen to each string carefully and type the words in the correct order when a text box appeared. Participants were asked to use correct spelling and separate each word with a space.

The study utilized Qualtrics survey software. Each participant was given the same task order and pseudo-randomized item order within each task, to reduce inter-individual variability due to order effects (James, Fraundorf, Lee, & Watson, 2018). Participation lasted 35 min.

2.4. Results

All analyses and hypotheses were preregistered (https://aspredicted.org/f4x8b.pdf). All data and code are available at: https://osf.io/sj9cf/.

2.4.1. Artificial language statistical learning

2AFC performance was significantly above chance ($M = 0.68$, $SD = 0.13$, $Range = 0.41$—1; $t(64) = 11.24$, $p < .0001$, $d = 1.40$).

For SICR,[8] two performance measures were calculated: the total number of syllables recalled (which measures the general impact of SL on basic recall abilities) and the number of full trigrams recalled. Trigram recall measures how well participants acquired the specific words from the input in the target trials (e.g., whether they recall *tagalu* and/or *lomari* in *tagalulomari*), which can be compared against baseline working memory for the items in the same positions in the random trials (syllables 1 + 2 + 3 and/or 4 + 5 + 6). Linear mixed-effects models were run on the SICR data using the "lmerTest" package (Kuznetsova, Brockhoff, & Christensen, 2017) in R, version 4.0.2 (R Core Team, 2020), with item type (target/random) as a fixed effect, and subject and items as random effects.

For the total number of syllables recalled, participants performed significantly better on the target over the random items ($\chi^2(1) = 27.86$, $p < .0001$; difference estimate = −0.93, $SE = 0.15$, $z = -6.39$, $p < .0001$). For the total number of syllable trigrams recalled, participants performed significantly better on the target items ($\chi^2(1) = 30.95$, $p < .0001$; difference estimate = −0.43, $SE = 0.06$, $z = -6.90$, $p < .0001$). The summary statistics for each SICR measure are reported in Table 1.

2.4.2. Natural language statistical learning

As with SICR, the total number of words correctly recalled was evaluated as a general measure of performance in MWC.[9] The number of word trigrams recalled was used to assess participants' sensitivity to specific multiword units. Linear mixed-effects models were constructed using item type (target/random) as a fixed effect, with subject and items as random effects.

For the total number of words recalled, participants performed significantly better on the target over the random items ($\chi^2(1) = 52.77$, $p < .0001$; difference estimate = −5.13, $SE = 0.33$, $z = -15.63$, $p < .0001$). Participants also recalled significantly more trigrams in the target items ($\chi^2(1) = 53.07$, $p < .0001$; difference estimate = −2.18, $SE = 0.14$, $z = -15.71$, $p < .0001$). The summary statistics for each MWC measure are reported in Table 2.

2.4.3. Correlations between tasks

Because we were interested in the degree to which sensitivity to trigrams in SL is associated with sensitivity to similar structures in natural language,[10] we conducted a series of correlational analyses using the SICR and MWC trigram difference scores[11] (target—random), which control for baseline working memory (recall of the random items). As predicted, SICR and MWC trigram recall were significantly correlated: r

Table 1
SICR results.

	% Syllables Recalled			% Syllable Trigrams Recalled		
	Mean	SD	Range	Mean	SD	Range
Target	65	18	31—97	44	25	3—92
Random	50	18	25—99	23	22	0—97

[6] Each syllable in the artificial language tasks and each word in MWC were synthesized individually, then combined with 75 ms pauses between each to create the input and test items. This was done in order to eliminate prosody and coarticulation.

[7] All stimuli are available at: https://osf.io/sj9cf/.

[8] Prior to analysis of the SICR data, an anchoring procedure was used to align syllable productions as closely as possible to the presented stimulus, to award maximal points for every syllable correctly recalled. Consistent syllable mistranscriptions were also corrected. See the Supplemental Materials for further information.

[9] Like SICR, an anchoring procedure was used for MWC, to grant participants maximum credit for every correctly recalled word. Misspellings were also corrected. The full details of the data pre-processing are reported in the Supplemental Materials.

[10] We solely analyzed the trigram scores, as these provide the most direct signature of sensitivity to the targeted statistical structure. By contrast, the total scores likely include sensitivity to other kinds of information, such as positional or bigram information.

[11] The difference scores are conservative estimates (Draheim, Mashburn, Martin, & Engle, 2019), as the maximum correlation of differences scores is limited by the error of both measures (target item recall and random item recall; Caruso, 2004), and difference scores amplify the amount of noise in the data (Zumbo, 1999).

Table 2
Multiword chunking results.

	% words recalled			% word trigrams recalled		
	Mean	SD	Range	Mean	SD	Range
Target	76	14	45—99	65	18	30—98
Random	34	17	15—94	11	18	0—85

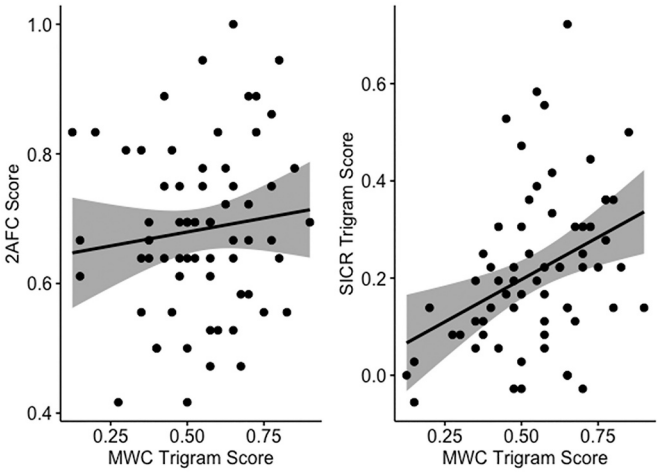

Fig. 1. Correlations between the artificial statistical learning measures and natural language statistical learning. SICR significantly correlates with sensitivity to natural language trigrams, whereas 2AFC does not.

(63) = 0.37, p = .002, whereas 2AFC[12] and MWC were not: $r(63) = 0.11$, p = .36 (Fig. 1). However, the two SL tasks, 2AFC and SICR trigram recall, were significantly correlated: $r(63) = 0.41$, p = .0008.

3. Discussion

Does sensitivity to statistical dependencies in the laboratory predict natural language abilities? Here, we took a step toward forging this link by aligning the specific statistical regularities involved in tasks of both kinds, and thus, presumably, the computations involved in detecting them. We found that SL in the lab predicts distributional sensitivity in the real world, but that differences arise due to task.

In our experiment, the classic 2AFC task did not correlate with natural language SL as measured by MWC. This may in part stem from the limited reliability of 2AFC both in our dataset and others (e.g., Arnon, 2020; Isbilen et al., 2020; Kidd et al., 2020). And although SICR and MWC have similar task demands in that they are both memory recall tasks, 2AFC also involves the same memory component as SICR. Both 2AFC and SICR require participants to keep two trisyllabic nonsense words in memory before making a response (a Yes/No key press for 2AFC and recalling the 6 syllables for SICR). Furthermore, previous individual differences work with MWC (McCauley et al., 2017) demonstrates that MWC does not correlate with nonword repetition (NWR) performance—the recall task that SICR was modeled on. It is thus not the case that recall tasks automatically correlate with one another by virtue of shared task demands. The NWR task in McCauley et al. (2017) manipulated the word-likeness of the stimuli, such that they were either similar or dissimilar to phoneme combinations in English. However, these phoneme combinations were not limited to trigram statistics, and thus tested broader aspects of statistical knowledge than MWC. Indeed, one might expect that the NWR and MWC tasks in McCauley et al. (2017) should show a stronger correlation than the one observed here: both tested sensitivity to natural language statistics at different levels of linguistic abstraction. By contrast, the artificial language in the current paper was designed to resemble English words as little as possible (see Isbilen et al., 2020, for further details). This further underscores the idea that the shared statistical computations across SICR and MWC—the processing of statistically-defined trigrams—most likely drove the correlation between the two. Thus, the shared recall component alone cannot explain the connection between SICR and MWC—not least because the SICR-MWC correlations control for baseline working memory—though we acknowledge that it may play a partial role.

Moreover, it is worth noting that while SICR lacks semantics, MWC inevitably draws upon stimuli with known meanings, and likely taps into aspects of syntactic knowledge. Despite this added complexity, and the fact that MWC involves processes beyond statistical computation alone, the correlation remains strong. Furthermore, semantics play a considerably smaller role in our task than the tests of reading and vocabulary typically used for comparisons with SL, where meaning is key to performance. Indeed, previous studies have revealed mixed correlations between SL and other natural language tasks involving semantic and syntactic processing, including reading and grammar (e.g., Gabay, Thiessen, & Holt, 2015; Haebig et al., 2017). If semantics and syntax were the only factors driving the effect in the SL-MWC correlation, then it stands to reason that SL should reliably correlate with these other natural language tasks that also rely on similar processes. Furthermore, while meaningfulness does play a role in the processing of multiword chunks (Jolsvai et al., 2020), there is a growing body of work showing that frequency has a strong effect on both the processing (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008) and production (e.g., Arnon & Clark, 2011; Arnon & Cohen-Priva, 2013) of multiword sequences (see Arnon & Christiansen, 2017, for a review).

Sensitivity to syllable trigrams derived from the Saffran et al. (1996) SL study as measured by SICR significantly predicted sensitivity to high-frequency word trigrams derived from natural language (multiword units being a building block for numerous linguistic skills). Short-term statistical learning in the lab induced changes to memory comparable to those produced by long-term learning in the real world, whereby participants display superior recall of statistical sequences that occur frequently in the environment (Baddeley, Conrad, & Hull, 1965; Jones & Macken, 2015; Miller & Selfridge, 1951). This suggests that SL taps into key aspects of natural language acquisition, with proficiency in assimilating novel statistics predicting proficiency in acquiring comparable structures from speech and written text. It also suggests that a general ability to discover adjacent statistical patterns may be common across different levels of linguistic abstraction, spanning individual words to multiword patterns. The inconsistent pattern of correlations between SL and other aspects of cognition may thus in part arise from an incongruence in the targeted statistical structures, in line with prior evidence pointing to the specificity of SL computations (e.g., Siegelman & Frost, 2015).

Usage-based theories have long advocated the centrality of multiword chunks to language, both in childhood and adulthood (e.g., Goldberg, 2006; Lieven, 2016). By detecting and storing statistically-contiguous multiword patterns in speech, individuals can abstract over encountered sequences to form novel generalizations, setting the stage for grammatical development and linguistic productivity—abilities that many SL studies aim to capture individual differences in. Usage of multiword chunks is a key signature of linguistic fluency, with second language learners producing significantly fewer multiword units than first language learners in speech and writing (Paquot & Granger, 2012). High-frequency multiword units are comprehended and produced faster than low-frequency units (Bannard & Matthews, 2008), similar to SL

[12] We computed the reliability/internal consistency of each measure (Cronbach's alpha; α). While SICR demonstrated excellent reliability (α = 0.95 for both the total and trigram scores), and MWC demonstrated acceptable to excellent reliability (total score: α = 0.75; trigram score = α = 0.90), 2AFC demonstrated questionable reliability (α = 0.68; see Supplemental Material for full details). The low reliability of 2AFC may have been one factor that influenced the uneven pattern of correlations.

reaction time data showing that participants are faster to respond to acquired statistical patterns (e.g., Hunt & Aslin, 2001). High frequency multiword chunks from natural speech are more robust to production errors (Arnon & Clark, 2011), similar to how production errors in SL tasks tend to occur at item boundaries where statistical probabilities are lower, rather than within statistically-coherent units (Krishnan, Carey, Dick, & Pearce, 2021). Here we provide behavioral evidence that the acquisition of multiword chunks may be underpinned by basic statistical learning mechanisms, supplementing recent computational modeling illustrating how distributional learning and memory processes work together to discover, produce, and comprehend language across 29 Old World languages (McCauley & Christiansen, 2019a).

Understanding the relationship between individual differences in SL and language is a longstanding goal of cognitive science. Here, we demonstrate that participants' proficiency in acquiring trigram structure in an artificial language significantly correlates with their sensitivity to high-frequency word trigrams in natural language. Designing studies that better tap into specific statistical knowledge and computations may thus provide an important steppingstone to interpreting the connection between SL and cognition at large.

Acknowledgements

This work was in part supported by the NSF GRFP (#DGE-1650441) and a Cornell Department of Psychology grant awarded to ESI.

Credit author statement

ESI, SMM, and MHC designed the study. ESI and SMM created the stimuli, and ESI performed the data analysis. All authors contributed to the writing of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105123.

Appendix A. SICR items

Target	Corresponding foil
kibudulatibi	tidubibulaki
kibudutopoka	bukapodukito
latibilomari	rilobimatila
latibitagalu	tabigatilula
lomarikibudu	lobukimaduri
lomarimodipa	moripadimalo
modipakibudu	dibumokidupa
modipatopoka	popamokadito
tagalulomari	tarimalugalo
tagalumodipa	gaditamolupa
topokalatibi	bikatolapoti
topokatagalu	kalutotapoga
lomaritopoka	tomakaloripo
modipatagalu	taludigamopa
kibudulomari	dumabulokiri
modipalatibi	patilamobidi
topokakibudu	kipobutokadu
tagalulatibi	tatigabilalu

Appendix B. MWC items

Target	Corresponding foil
had a dream kind of silly something to say on a diet	diet a silly say dream kind to of a something on had
have a secret time to stop all the hype in the mailbox	secret the have all to mailbox time the a in hype stop
to the edge have some fun don't know me not really familiar	familiar the to fun have know edge not don't some me really
don't like them good to know is really nice have to eat	really them nice have eat know to don't good like is to
at the moment what you said one of these I guess not	of I one what guess said moment at not the you these
take the quiz sample of data really don't matter this is typical	is take data really the this don't of sample quiz matter typical
such a burden that's the agreement into the unknown nothing to wear	unknown wear into the that's to such the nothing agreement burden a
a personal nature get a certificate across the highway take a stroll	highway a take personal the across a nature a get stroll certificate
off the path see the picture kind of disturbing a bad attitude	attitude a off disturbing the see bad of picture kind path the
its a lie when I die be a burden in a dispute	when burden lie a be I a its die a dispute in

References

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*, 286–304.

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods, 52*, 68–81. https://doi.org/10.3758/s13428-019-01205-5

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science, 9*, 621–636.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth–Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development, 7*, 107–129.

Arnon, I., & Cohen-Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*, 349–371. https://doi.org/10.1177/0023830913484891

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*, 67–82.
Baddeley, A. D., Conrad, R., & Hull, A. J. (1965). Predictability and immediate memory for consonant sequences. *Quarterly Journal of Experimental Psychology, 17*, 175–177.
Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*, 241–248.
Bogaerts, L., Siegelman, N., Christiansen, M. H., & Frost, R. (2022). Is there such a thing as a "good statistical learner?". *Trends in Cognitive Science, 26*, 25–37.
Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment, 20*, 166–171.
Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science, 9*, 542–551.
Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2005). *Fisher English training speech (part 2) transcripts*. Philadelphia: Linguistic Data Consortium.
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.
Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 24–39.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–114.
Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin, 145*, 508–535.
Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44.
Gabay, Y., Thiessen, E. D., & Holt, L. L. (2015). Impaired statistical learning in developmental dyslexia. *Journal of Speech, Language, and Hearing Research, 58*, 934–945.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Haebig, E., Saffran, J. R., & Ellis Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry, 58*, 1251–1263.
Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130*, 658–680.
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*, Article e12848.
James, A. N., Fraundorf, S. H., Lee, E. K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language, 102*, 155–181.
Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2020). Meaningfulness beats frequency in multiword chunk processing. *Cognitive Science, 44*, Article e12885.
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. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development, 87*, 184–193.
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*, Article 104964.
Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences, 22*, 154–169.
Krishnan, S., Carey, D., Dick, F., & Pearce, M. T. (2021). Effects of statistical learning in passive and active contexts on reproduction and recognition of auditory sequences. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001091. Advance online publication.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26. https://doi.org/10.18637/jss.v082.i13
Lieven, E. (2016). Usage-based approaches to language development: Where do we go from here? *Language and Cognition, 8*, 346–368.
Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24*, 187–219.
McCauley, S. M., Bannard, C., Theakston, A., Davis, M., Cameron-Faulkner, T., & Ambridge, B. (2021). Multiword units lead to errors of commission in children's spontaneous production: "What corpus data can tell us?". *Developmental Science, 24*, Article e13125. https://doi.org/10.1111/desc.13125
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.
McCauley, S. M., Isbilen, E. S., & Christiansen, M. H. (2017). Chunking ability shapes sentence processing at multiple levels of abstraction. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2681–2686). Austin, TX: Cognitive Science Society.
Miller, G. A., & Selfridge, J. A. (1951). Verbal context and the recall of meaningful material. *American Journal of Psychology, 63*, 176–185.
Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62*, 302–331.
Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130–149.
Qi, Z., Sanchez Araujo, Y., Georgan, W., Gabrieli, J., & Arciuli, J. (2019). Hearing matters more than seeing: A cross- modality study of statistical learning and reading ability. *Scientific Studies of Reading, 23*, 101–115.
R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/.
Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC): Second rselease*. Philadelphia: Linguistic Data Consortium.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.
Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120.
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 Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2021). The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia. *Dyslexia, 27*, 168–186.
Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (pp. 269–304). Greenwich: JAI Press.