Isbilen, E.S., Frost, R.L.A, Monaghan, P. & Christiansen, M.H. (2018). Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning. In T.T. Rogers, M. Rau, X. Zhu, & C.W. Kalish (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 1856-1861). Austin, TX: Cognitive Science Society.

Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning

Erin S. Isbilen (esi6@cornell.edu)
Cornell University, Department of Psychology, Ithaca, NY 14850 USA

Rebecca L.A. Frost (rebecca.frost@mpi.nl)
Max Planck Institute for Psycholinguistics, Language Development Department, Nijmegen, 6525 XD, Netherlands

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Lancaster University, Department of Psychology, Lancaster, LA1 4YF, UK

Morten H. Christiansen (christiansen@cornell.edu)
Cornell University, Department of Psychology, Ithaca, NY 14850 USA

Abstract

A common assumption in the cognitive sciences is that artificial and natural language learning rely on shared mechanisms. However, attempts to bridge the two have yielded ambiguous results. We suggest that an empirical disconnect between the computations employed during learning and the methods employed at test may explain these mixed results. Further, we propose statistically-based chunking as a potential computational link between artificial and natural language learning. We compare the acquisition of non-adjacent dependencies to that of natural language structure using two types of tasks: reflection-based 2AFC measures, and processing-based recall measures, the latter being more computationally analogous to the processes used during language acquisition. Our results demonstrate that task-type significantly influences the correlations observed between artificial and natural language acquisition, with reflection-based and processing-based measures correlating within – but not across – task-type. These findings have fundamental implications for artificial-to-natural language comparisons, both methodologically and theoretically.

Keywords: statistical learning; chunking; language; artificial language learning; cross-situational learning; non-adjacent dependencies; learning; memory; serial recall; methodology

Introduction

Connecting individual differences in artificial and natural language learning is an ongoing endeavor in the cognitive sciences. These studies operate on the assumption that artificial language learning tasks designed for use in the laboratory draw on the same cognitive processes that underpin language acquisition in the real world (e.g., Saffran, Aslin, & Newport, 1996). Yet, attempts to bridge artificial and natural language learning have yielded mixed results, often finding weaker correlations between language measures that should in theory rely on shared computations (Siegelman, Bogaerts, Christiansen & Frost, 2017). Part of the problem may lie in the nature of the tests used to evaluate learning; although artificial and natural language learning may rely on the same computational processes, different tests may tap into separate subcomponents of these skills, making the relationship difficult to unpack.

Artificial language learning tasks are assumed to capture key aspects of how learners acquire language in the real world: by drawing on the distributional information contained in speech. Through exposure to statistical regularities in the input, the cognitive system picks up on linguistic units without awareness by the learner (Saffran et al., 1996). Yet, in adults, statistical learning is typically tested using measures that require participants to reflect on their knowledge and provide an overt judgment, such as in the two-alternative forced-choice task (2AFC); a test that, while potentially informative, only provides a meta-cognitive measure of learning. Indeed, language learning measures can be broadly divided into two categories: *reflection-based* measures (e.g., 2AFC), which translate the primary effects of learning into a secondary response, and *processing-based* measures, which rely on the same computations as the learning itself (Christiansen, 2018). In psycholinguistic research, it is often the case that the learning measures employed at test do not align with the processes employed during learning. We propose that this disconnect may have constrained prior observations of the relationship between artificial and natural language skills. We seek to resolve some of this ambiguity in the study at hand.

In the current study, we assess the degree to which statistical learning abilities map onto natural language acquisition, and evaluate correlations within and between reflection- and processing-based measures. For the purpose of this paper, we characterize artificial language learning as statistical learning in a highly constrained, simplified context, using the Saffran et al. (1996) familiarization method. We simulate natural language acquisition by presenting participants with a more complex cross-situational learning task that utilizes natural vocabulary and grammar with corresponding referents. For each part of the experiment, we included two types of tests: reflection-based tasks (2AFC), and processing-based tasks (recall), to allow for a comparison of learning between and within task types.

For our processing-based measure, we employed a chunking-based recall task, building on the suggestion that chunking plays a key role in statistical learning and language acquisition (see Christiansen, 2018, for a review).

In 2AFC tasks, participants are required to indicate their preference for one stimulus over another, which is taken to indicate learning. In recall – a task which is thought to rely on chunking – participants repeat syllable strings that are either congruent or incongruent with the statistics of the input, with recall errors acting as a window into learning. That is, learning is indexed by better recall of consistent items when controlling for baseline phonological working memory (Conway, Bauernschmidt, Huang & Pisoni, 2010; Isbilen, McCauley, Kidd & Christiansen, 2017). If chunking occurs during language acquisition, chunking-based tasks may yield a better measure of learning than reflection-based tasks such as 2AFC.

In the first part of the experiment, participants engaged in a statistical learning task adapted from Frost and Monaghan (2016), to test segmentation and generalization of non-adjacent dependencies (the artificial language task). In the second part, participants learned a fragment of Japanese, comprising a small vocabulary and simple grammar using a cross-situational learning task adapted from Rebuschat, Ferrimand, and Monaghan (2017) and Walker, Schoetensack, Monaghan and Rebuschat (2017; the natural language task). We hypothesized that the correlations observed between artificial and natural language learning would show a strong effect of task type: reflection-based measures would be more likely to correlate with other reflection-based measures, whereas processing-based measures would be more likely to correlate with other processing-based measures. Such a pattern would have important implications for individual differences work, and about the deductions that can be applied to natural language acquisition from artificial language learning tasks.

Part 1: Non-adjacent dependency learning in an artificial language

In Part 1, we tested adults' learning of an artificial language composed of non-adjacent dependencies, which are relationships between linguistic units that occur across one or more variable intervening units (e.g., in an AXC structure where units A and C reliably co-occur, but X varies independently). These dependencies are found at multiple levels of linguistic abstraction, including morphology within words and syntactic dependencies between words, thereby providing a tightly-controlled artificial structure that shares structural similarity with natural language.

We examined learners' ability to segment these non-adjacent dependency sequences from speech, and generalize them to new instances - skills which are integral to natural language learning. We tested both segmentation and generalization with a reflection-based task (2AFC), and a processing-based task, the statistically-induced chunking recall task (SICR; Isbilen et al., 2017). In the SICR task, participants are presented with 6-syllable-long strings, that are either composed of two words from the input, or the same syllables presented in a random order. If participants have successfully chunked the items in the artificial language during training, we expect that they should perform significantly better on recalling the strings derived from the statistics of the input language. While 2AFC is scored as a correct-incorrect binary, SICR is scored syllable-by-syllable, which we suggest may provide more in-depth insights into segmentation and generalization skills. Building on the results of Frost and Monaghan (2016), we hypothesized that both tasks would yield evidence of simultaneous segmentation and generalization. However, due to the differences in task demands between reflection- and processing-based tests, we expected to see limited correlations between measurement types.

Method

Participants 49 Cornell University undergraduates (30 females; age: *M*=19.43, *SD*=1.30) participated for course credit. All participants were native English speakers, with no experience learning Japanese.

Materials The same language and stimuli as Frost and Monaghan (2016) were used, derived from Peña, Bonatti, Nespor and Mehler (2002). The language was composed of 9 consonant-vowel syllables (*be, du, fo, ga, li, ki, pu, ra, ta*), arranged into three tri-syllabic non-adjacent dependencies containing three varying middle syllables ($A_1X_{1-3}C_1$, $A_2X_{1-3}C_2$, and $A_3X_{1-3}C_3$; 9 words in total). Four different versions of the language were created to control for potential preferences for certain phoneme combinations. Syllables used for the A and C items contained plosives (*be, du, ga, ki, pu, ta*), while the X syllables contained continuants (*fo, li, ra*). The resulting 9 items are referred to as *segmentation words*, sequences that were presented during training. Nine *generalization words* were also created, and were only presented at test. The generalization words contained trained non-adjacent dependencies, but with novel intervening syllables (*thi, ve, zo*, e.g., $A_1Y_{1-3}C_1$). The generalization words measure participants' ability to extrapolate the knowledge of the non-adjacent dependencies gained during training to novel, unheard items.

For the 2AFC test, 18 additional foil words were created, which were paired with segmentation and generalization words. Foils for the segmentation test comprised part-word sequences that spanned word boundaries (e.g., CAX, XCA). Foils for the generalization test were part-words but with one syllable switched out and replaced with a novel syllable, to prevent participants from responding based on novelty alone (e.g., NCA, XNA, CAN, see Frost & Monaghan, 2016). For the SICR test, 27 six-syllable strings were created: 9 composed of two concatenated segmentation words (e.g., $A_1X_1C_1A_2X_2C_2$), 9 composed of two generalization words (e.g., $A_1Y_1C_1A_2Y_2C_2$), and 9 foils. The foils used the same syllables as the experimental items in a pseudorandomized order that avoided using any

transitional probabilities or non-adjacent dependencies from the experimental items.

All stimuli were created using the Festival speech synthesizer (Black et al., 1990). Each AXC string lasted ~700 ms, and was presented using E-Prime 2.0.

Procedure For training, the 9 segmentation words were concatenated into a continuous stream that participants heard for 10.5 minutes. Participants were instructed to listen carefully to the language and pay attention to the words it might contain.

To test learning, two different tasks were used: the 2AFC task and the SICR task (Isbilen et al., 2017). The order of the two tests was counterbalanced to account for potential order effects. In the 2AFC task, participants were presented with 18 pairs of words: 9 segmentation pairs and 9 generalization pairs, with each pair featuring a target word and corresponding foil. Segmentation and generalization trials were randomized within the same block of testing. Participants were instructed to carefully listen to each word pair and indicate which of the two best matched the language they heard during training. In the SICR task, 27 strings were randomly presented for recall: 9 segmentation items, 9 generalization items, and 9 foils that served as a baseline working memory measure. Participants were asked to listen to each string and say the entire string out loud to the best of their ability. Participants were not informed of any underlying structure of the strings in either task.

Results and Discussion

First, we examined the data for task order effects (2AFC first/SICR second versus SICR first/2AFC second), and language effects (which of the four randomized languages participants heard). A one-way ANOVA revealed a significant effect of order on both SICR measures (Segmentation: $F(3,45)=-2.30$, $p=.026$; Generalization: $F(3,45)=-3.30$, $p=.002$), with participants who received 2AFC prior to SICR scoring significantly higher on these two measures. Similarly, language significantly impacted SICR generalization performance, $F(3,45)=6.94$, $p=.0006$, suggesting that different syllable combinations may vary in difficulty when being spoken aloud. All subsequent analyses involving SICR in the remainder of the paper control for order, and for SICR generalization, for both order and language.

2AFC Performance Replicating the findings of Frost and Monaghan (2016), participants showed simultaneous segmentation and generalization of non-adjacent dependencies, with performance on both tasks being significantly above chance (Segmentation: $M=.84$, $SD=.13$; $t(48)= 18.44$, $p<.0001$; Generalization: $M=.70$, $SD=.21$; $t(48)= 6.61$, $p<.0001$). Performance was significantly more accurate on segmentation than generalization trials, $t(48)=5.77$, $p<.0001$, and segmentation and generalization scores were highly correlated: $r(47)=.53$, $p<.0001$.

SICR Performance Participants' verbal responses from the SICR task were transcribed by two coders blind to the study design. The transcriptions were subsequently scored against the target test items to obtain measures of overall accuracy (the total number of syllables recalled in the correct serial position), and non-adjacent dependency accuracy (the number of A-C pairings recalled from each item, out of the two possible pairings: e.g., $\underline{A_1}x\underline{C_1}\underline{A_2}x\underline{C_2}$). Replicating the results of Isbilen et al. (2017), participants accurately recalled significantly more syllables in the correct order for the experimental items than the random items. These results held for both the segmentation items (Experimental: $M=34.84$, $SD=10.16$; Random: $M=13.55$, $SD=6.04$; $t(48)= 23.11$, $p<.0001$), and for the generalization items (Experimental: $M=27.71$, $SD=10.56$; Random: $M=8.35$, $SD=4.37$; $t(48)= 15.98$, $p<.0001$). Similarly, the number of non-adjacent dependencies (syllables in the 1st & 3rd and/or the 4th & 6th serial positions) recalled for experimental items was significantly higher than those recalled for the random, both for the segmentation items (Experimental: $M=8.53$, $SD=4.52$; Random: $M=2.57$, $SD=2.34$; $t(48)= 13.34$, $p<.0001$), and for the generalization items (Experimental: $M=6.63$, $SD=4.42$; Random: $M=1.50$, $SD=1.45$; $t(48)= 9.51$, $p<.0001$). Unlike 2AFC, the SICR results revealed no significant difference in performance between the segmentation and generalization difference scores (experimental minus random), although generalization scores were slightly lower due to the inclusion of unfamiliar syllables. These results held for both overall recall, and for the total number of non-adjacent dependencies recalled ($p=.10$ in both cases). This difference between 2AFC and SICR may in part stem from differences in task demands: differences in familiarity between segmentation and generalization items may influence ratings in 2AFC more, due to its meta-cognitive nature. SICR generalization and segmentation performance was significantly correlated: $r(47)=.34$, $p=.02$.

Correlations between 2AFC and SICR To evaluate the relationship between reflection- and processing-based measures, correlations were run between 2AFC and SICR scores. The SICR values used for the correlations were the total difference scores, to maximize the measures' comparability to 2AFC (which is akin to a difference score), while also controlling for baseline phonological working memory (the subtraction of the random items from the experimental items). The only significant correlation was between 2AFC segmentation and SICR segmentation, $r(47)=.31$, $p=.04$ (not correcting for multiple comparisons).[1] No other SICR and 2AFC measures were significantly correlated (all $p>.08$). In line with our hypothesis, these findings suggest that reflection- and processing-based measures appear to capture largely different aspects of statistical learning skills (Christiansen, 2018; Siegelman et al., 2017).

The results of Part 1 replicate the results of Frost & Monaghan (2016) of simultaneous segmentation and generalization of non-adjacent dependencies using both 2AFC and SICR. Taken together, these findings suggest that

[1] In a pilot version of this study, (N=61) no such correlation was observed, potentially suggesting a type II error: $r(59)=-.04$, $p=.76$.

statistical-chunking processes may be able to account for the segmentation and generalization of non-adjacent dependencies, as well as that of sequential dependencies. Furthermore, we found that although reflection- and processing-based measures showed evidence of learning, performance across the two tasks was largely uncorrelated. To test whether this pattern extends to natural language acquisition, Part 2 of the experiment evaluated grammar and vocabulary acquisition using patterns from natural language, with a comparison of 2AFC and recall task types.

Part 2: Cross-situational language learning of Japanese

Natural language acquisition involves a host of different factors, including word segmentation, word-referent mapping, discovery of sequential structure, and generalization to novel instances. In the second part of this experiment, we increased the complexity of the task to explore the degree to which the learning taking place during segmentation and the discovery of non-adjacent dependency structure maps onto more naturalistic language acquisition. A cross-situational language learning task based on Walker et al. (2017) was administered, exposing participants to Japanese sentences co-occurring with complex scenes. Cross-situational learning simulates naturalistic language learning in the lab by analogy to infants' acquisition of word-referent mappings non-ostensively, through hearing instances of a word occurring with the same referent across different contexts. Similar to Part 1, both reflection-based and processing-based measures were used to evaluate learning. 2AFC tests of noun, marker, and verb learning were performed. Additionally, a combined forced-choice and recall task was also administered to test syntax acquisition: participants repeated whole sentences they heard out loud, after which they rated the grammaticality of each sentence. We hypothesized that vocabulary and grammatical regularities would be acquired simultaneously, similar to the concurrent segmentation and generalization in the non-adjacent dependency task. Furthermore, we anticipated that while all tests would show some evidence of learning, only within task-type correlations would be significantly related.

Method

Participants The same 49 participants from Part 1 partook in Part 2 immediately following the first task.

Materials A small lexicon of Japanese words was used for this experiment, taken from Rebuschat et al. (2017). The language consisted of 6 nouns (*fukuoru*, owl; *kame*, turtle; *niwatori*, chicken; *shimauma*, zebra; *ushi*, cow; *zou*, elephant), four verbs (*kakusu*, hide; *mochiageru*, lift; *taosu*, push; *tobikueru*, jump), and two case markers (*-ga* and *-o*), which were appended to the end of each noun to indicate whether the noun was the subject (*-ga*) or the object (*-o*) of the sentence. For instance, the sentence "*kamega shimaumao taosu*" would indicate that the turtle (subject) pushes the zebra (object). The language also used Japanese syntax, with sentences having two possible grammatical orders: subject-object-verb (SOV), and object-subject-verb (OSV). For training, 192 sentences were generated. For test, 96 additional sentences were presented: 24 for each of the marker, noun, and verb tests, and 24 for the combined syntax and recall task. Of the syntax stimuli, 12 were ungrammatical items that used word orderings that are invalid according to Japanese syntax (OVS, VOS, VSO, SVO). The frequency, order, and object-subject assignment of each word were all balanced. All auditory training and test stimuli were created by a native Japanese speaker.

With each sentence, complex scenes depicting cartoon animals as the referents for the nouns were also presented, engaging in the action indicated by the verb of each sentence (hiding, lifting, pushing, or jumping). During training, two such scenes were presented, one the target scene and the other a distractor, to allow for the accrual of word-referent mappings through the use of cross situational statistics. During the syntax test, only the target scene was presented. All stimuli were presented in E-Prime 2.0.

Procedure The experiment consisted of two training blocks, and two test blocks. During training participants heard a Japanese sentence while watching two scenes play on the computer screen: one displaying the target, and the other the foil. The foil scene varied from the target both in terms of the nouns and actions depicted. Participants were asked to judge to the best of their ability which scene the sentence referred to. Unknown to the participant, the last trials of training tested their knowledge of the nouns, verbs, and markers of the language, using the same method by varying the two scenes by just one property (e.g., only one object was different, or only the action was different, or only the subject/object roles were different). Following training, 12 syntax test trials were administered, which presented a sentence paired with a single scene. Participants were told that the speaker of these sentences were learning Japanese, and that their task was to repeat the speaker's sentence out loud (the recall measure), and then indicate whether the speaker's sentence sounded "good" or "funny" (the forced-choice measure). While this task is slightly different from the other 2AFC measures in this experiment, in which participants choose between a target and a foil, they both require participants to engage in reflection about learned material.

After the conclusion of the first training block and syntax test, the same procedure was completed once more, starting with training and ending with another syntax test. Each training block contained 4 marker test trials, 4 verb test trials, and 6 noun test trials. Each syntax test block contained 12 test trials: 6 grammatical, and 6 ungrammatical sentences. No feedback about participants' performance was provided at any time during training or test.

Results and Discussion

2AFC results Following the methods employed by Walker et al. (2017), data from both testing blocks were pooled for the analyses. With the exception of noun learning *(M*=.54,

SD=.20; *t*(48)=1.58, *p*=.12), scores on all 2AFC measures were significantly above chance (Marker test: *M*=.57, *SD*=.19; *t*(48)=2.80, *p*=.0075; Verb test: *M*=.60, *SD*=.14; *t*(48)=5.52, *p*=.0075; Syntax test: *M*=.61, *SD*=.14; *t*(48)=5.52, *p*=.0075). Thus, our results showed that vocabulary and grammatical acquisition of natural language structure can occur simultaneously. The lack of significant noun learning, while inconsistent with the findings of Walker et al. (2017), may be explained by the fact that the nouns in this task were longer (containing more syllables), which may have contributed to reduced learning. The only 2AFC tests that were significantly correlated with each other were performance on the syntax and verb test, *r*(47)=.39, *p*=.0065, which is consistent with the findings of Walker et al. (2017).

Syntax recall results Participants' verbal responses were transcribed by a coder blind to the study design, and were scored against the targets, with a point given for each syllable recalled in the correct serial position. Overall, no effect of grammaticality was found on recall performance, with participants recalling approximately equal numbers of syllables in the correct order for both the grammatical (*M*=73.08, *SD*=19.46) and ungrammatical items (*M*=72.14, *SD*=18.57; *t*(48)= .60, *p*=.55). However, unlike the artificial language stimuli, the natural language recall items vary in the total number of syllables, ranging from 8 to 14 syllables. A linear mixed effects model of the raw recall data revealed that while grammaticality and string length had no effect independently on recall performance, the interaction of grammaticality and string length was significant, *t*(1147)=2.78, *p*=.0055, with length relating to significant detriments in recall scores for ungrammatical items, but not grammatical ones. This suggests that learning the statistical structure of the language stabilized recall of the grammatical items independent of length, whereas the ungrammatical items were more severely impacted by memory limitations.

Correlations between 2AFC and recall To investigate the connection between the reflection- and processing-based measures in the cross-situational learning task, correlations were run between all 2AFC measures in Part 2 of the experiment, and the recall difference scores. No significant correlations were found between any of the 2AFC measures and recall performance (all *p*>.14). These results may be taken as further evidence that reflection- and processing-based tests do not measure the same aspects of learning.

The relationship between artificial & natural language learning

To determine the connection between non-adjacent dependency learning in an artificial context to vocabulary and grammar acquisition in a natural language context, correlations were calculated between the data from Parts 1 and 2 of the experiment. These analyses were first performed within task type, then between task types. We predicted that the 2AFC measures from Part 1 would only correlate with the 2AFC measures from Part 2, and that only the recall measures from Parts 1 and 2 would be correlated.

Parts 1 & 2: Reflection-based measures

Correlations between all reflection-based (2AFC) measures were performed. However, only two 2AFC measures were correlated across the two parts of the experiment. First, participants' ability to segment words in the non-adjacent dependency task positively correlated with their ability to learn the nouns in the cross-situational learning task, *r*(47)=.35, *p*=.0139. Second, generalization ability on the non-adjacent dependency 2AFC task negatively correlated with participants' ability to pick up on the markers on the cross-situational learning task, *r*(47)=-.28, *p*=.0496.

Parts 1 & 2: Processing-based measures

Partial correlations between the SICR and cross-situational recall performance raw data, controlling for string length, item type (experimental versus random for SICR, or grammatical versus ungrammatical for the cross-situational items) and repeated measures revealed that SICR and cross-situational recall abilities significantly correlated with one another. SICR segmentation (*r*(615)=.20, *p*<.0001) demonstrated a stronger correlation to cross-situational recall than did SICR generalization (*r*(561)=.14, *p*<.0009).

Parts 1 & 2: Between measure-types

Correlations between reflection- and processing-based tests from Parts 1 and 2 of the experiment were performed. The results revealed no significant correlations between task types from the two parts of the experiment (Table 1).

Table 1: 2AFC and Recall Correlations

	Marker Test	Noun Test	Syntax Test	Verb Test	2AFC Seg.	2AFC Gen.
SICR Seg.	.02	.02	.12	.13	.31*	.26
SICR Gen.	-.09	.08	-.07	.15	.12	.16
Cross-sit. recall	-.22	.18	.09	-.02	.17	.15

General Discussion

Bridging artificial and natural language learning is an important endeavor in the cognitive sciences. A first step to stabilizing the link between in-lab observations and real-world behavior may come from strengthening the connection between the tasks used to test learning, and the computations employed during learning. Here, we argue for the role of statistically-based chunking as the computational link between learning and testing.

Short-term memory recall is a robust indicator of long-term learning (Jones & Macken, 2015; McCauley, Isbilen & Christiansen, 2017), with the accrual of statistical regularities over time aiding memory retention in the here-and-now. The use of such recall tasks as a measure of in-lab language learning is motivated by evidence supporting the notion that chunking may play a key role in statistical learning, and can account for language acquisition,

processing, and production more broadly (Christiansen & Chater, 2016). Our results strengthen this argument by demonstrating that statistically-based chunking can also account for the simultaneous learning and generalization of non-adjacent dependencies: a complex, dynamic linguistic structure. While our findings appear to support some degree of separability between segmentation and generalization skills (see also Frost & Monaghan, 2017), these abilities also appear to inform one another, with segmentation performance strongly predicting generalization ability.

While our study has important methodological and theoretical implications for individual differences work, it also has a number of limitations. First, although our cross-situational paradigm simulates aspects of natural language acquisition in the lab, it does not capture language acquisition exactly as it occurs in the real world. Second, while this language learning experiment implemented many separate subcomponents of natural language by design, we also acknowledge that these features changed the task demands. Learning in the artificial language task involved only segmentation and generalization of individual words, whereas the cross-situational learning task incorporated complex grammar, referents, and whole sentences. Stronger correlations may have been observed if the structure of the two different tasks were more similar (see Siegelman et al., 2017, for discussion).

Methodologically, our results suggest that the empirical disconnect between the learning targeted and the measures used at test may influence the correlations observed between artificial and natural language outcomes. While the processes leveraged for both in-lab and real-world language acquisition may be analogous, the similarity or dissimilarity of the tasks used to measure learning – and the specific computations each task relies on – may obscure the connection between the targeted cognitive processes. Moreover, there appears to be substantial individual variation in performance on these two kinds of tasks, as evidence of learning on one measure does not necessarily translate to high performance on the other. While both reflection- and processing-based measures test learning, our results suggest that they may test slightly different kinds of knowledge: meta-cognitive reflections over what was learned, versus processing-based facilitation from accrued statistics.

Acknowledgments

We would like to thank Phoebe Ilevebare, Farrah Mawani, Eleni Kohilakis, Olivia Wang, Dante Dahabreh, and Jake Kolenda for their help collecting and coding data. This work was in part supported by NSF GRFP (#DGE-1650441) awarded to ESI, a Cornell Department of Psychology Graduate Research Grant awarded to ESI, and 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].

References

Christiansen, M.H. (2018). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science*.

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

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.

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

Frost, R.L.A., & Monaghan, P. (2017). Sleep-driven computations in speech processing. *PloS one*, *12*(1), e0169538.

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

Isbilen, E.S., McCauley, S.M., Kidd, E. & Christiansen, M.H. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

McCauley, S.M. & Christiansen, M.H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. *Proceedings of the 33rd Annual Conference of the Cognitive Science*. 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. *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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

Rebuschat, P., Ferrimand, H., & Monaghan, P. (2017). Age effects in statistical learning of Japanese: Evidence from the cross-situational learning paradigm. Talk presented at the *International Association for the Study of Child Language*.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606-621.

Siegelman, N., Bogaerts, L., Christiansen, M. H. & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Phil. Trans. R. Soc. B*, *372*(1711), 20160059.

Walker, N., Schoetensack, C., Monaghan, P., & Rebuschat, P. Simultaneous acquisition of vocabulary and grammar in an artificial language learning task. *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.