Cognition 177 (2018) 198–213

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Linguistic entrenchment: Prior knowledge impacts statistical learning performance

Noam Siegelman[a,*], Louisa Bogaerts[a], Amit Elazar[a], Joanne Arciuli[b], Ram Frost[a,c,d]

[a] The Hebrew University of Jerusalem, Israel
[b] University of Sydney, Australia
[c] Haskins Laboratories, New Haven, CT, USA
[d] BCBL, Basque Center of Cognition, Brain and Language, San Sebastian, Spain

ARTICLE INFO

Keywords:
Statistical learning
Prior knowledge
Entrenchment
Domain generality vs. domain specificity

ABSTRACT

Statistical Learning (SL) is typically considered to be a domain-general mechanism by which cognitive systems discover the underlying statistical regularities in the input. Recent findings, however, show clear differences in processing regularities across modalities and stimuli as well as low correlations between performance on visual and auditory tasks. Why does a presumably domain-general mechanism show distinct patterns of modality and stimulus specificity? Here we claim that the key to this puzzle lies in the prior knowledge brought upon by learners to the learning task. Specifically, we argue that learners' already entrenched expectations about speech co-occurrences from their native language impacts what they learn from novel auditory verbal input. In contrast, learners are free of such entrenchment when processing sequences of visual material such as abstract shapes. We present evidence from three experiments supporting this hypothesis by showing that auditory-verbal tasks display distinct item-specific effects resulting in low correlations between test items. In contrast, non-verbal tasks – visual and auditory – show high correlations between items. Importantly, we also show that individual performance in visual and auditory SL tasks that do not implicate prior knowledge regarding co-occurrence of elements, is highly correlated. In a fourth experiment, we present further support for the entrenchment hypothesis by showing that the variance in performance between different stimuli in auditory-verbal statistical learning tasks can be traced back to their resemblance to participants' native language. We discuss the methodological and theoretical implications of these findings, focusing on models of domain generality/specificity of SL.

1. Introduction

The demonstration that infants can extract statistical properties from continuous speech (Saffran, Aslin, & Newport, 1996) has set the foundations for modern research on Statistical Learning (SL). The study by Saffran et al. (1996) offered a new perspective on how language is acquired by highlighting experience-based principles for detecting regularities in the environment, mainly, the tracking of transitional probabilities (TPs) between adjacent elements in sequentially presented input. In the many studies that followed, this initial demonstration was extended to different modalities (e.g., Fiser & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002), stimuli (e.g., Brady & Oliva, 2008; Gebhart, Newport, & Aslin, 2009), and ages (e.g., Arciuli & Simpson, 2011; Bulf, Johnson, & Valenza, 2011; Campbell, Zimerman, Healey, Lee, & Hasher, 2012), leading to the widespread perception that SL reflects domain-general cognitive computations for extracting and recovering the statistical regularities embedded in any sensory input (see Frost, Armstrong, Siegelman, & Christiansen, 2015, for a review).

At the core of this widely accepted view of SL is the assumption that there is something "common" underlying the learning of regularities across domains. Yet, a range of recent findings seem to challenge this assumption. First, domain-generality, as a theoretical construct, requires that at least some commonalities should exist in computing TPs across sets of visual and auditory stimuli, even if there are some inherent differences in perceiving regularities in different modalities. However, when this was tested by looking at correlations between individual performance across different SL tasks, the results consistently did not support domain-generality. For example, Siegelman and Frost (2015) reported that while the ability to extract TPs in the visual and auditory modality is a stable characteristic of the individual (with a test-retest reliability of around 0.6), correlation between performance in the auditory SL task (modeled on Saffran, Newport, & Aslin, 1996),

* Corresponding author at: Department of Psychology, The Hebrew University, Jerusalem 9190501, Israel.
E-mail address: noam.siegelman@gmail.com (N. Siegelman).

https://doi.org/10.1016/j.cognition.2018.04.011
Received 22 August 2017; Received in revised form 8 April 2018; Accepted 11 April 2018
0010-0277/ © 2018 Elsevier B.V. All rights reserved.

and a parallel task in the visual modality (modeled on Turk-Browne, Junge, & Scholl, 2005), is virtually zero.[1] Why is it that there is no trace of shared computations across modalities? Even more puzzling, Erickson and her colleagues have recently examined individual performance in two similar auditory SL tasks that varied only in their syllabic components (Erickson, Kaschak, Thiessen, & Berry, 2016). Similar to Siegelman and Frost (2015), they reported that performance for a given set of syllables was highly reliable, with a test-retest reliability spanning between 0.59 and 0.66. However, individual-level correlation in performing the two auditory SL tasks was strikingly low and not significant ($r = 0.17$).[2] Why is it that the seemingly random choice of "words" (i.e. the syllables that co-occur within a familiarization stream) leads to very different learning outcomes, when the same mechanism presumably computes the statistical properties of any speech stream?

A recent developmental study tracking visual and auditory SL performance at different ages (Raviv & Arnon, 2017) showed another puzzling outcome. Whereas visual SL performance improved linearly with age (7–12 years, and see Arciuli & Simpson, 2011, for similar findings), auditory SL performance, albeit lower on the average, did not show any improvement with age. If there is something like a domain-general mechanism for extracting patterns across modalities, why do we observe different developmental trajectories in the visual and auditory modalities?

Another puzzle concerns the very different results obtained with identical auditory SL tasks across speakers of different languages. Two recent studies, one with Italian speakers and one with French speakers, employed an identical experimental design to compare performance on "words" and "phantom words" (sequences of syllables that have the same TP structure as "words" but that never occur in the familiarization steam as a chunk). Surprisingly, these two studies found a virtually opposite pattern of results: In the study with Italian speakers, Endress and Mehler (2009) found that participants were equally familiar with "words" and "phantom words", and concluded that "phantom words" are treated as words. In contrast, in the study with French speakers (Perruchet & Poulin-Charronnat, 2012) consistent preference for "words" over "phantom words" was observed, which suggests that phantoms are not treated as words but rather as non-words. Since the experience-based principles for detecting regularities in continuous speech are supposedly universal, and certainly not privileged to the speakers of only a subset of natural languages, why is it that the language background of the participants appears to determine the outcome of the study?

What is going on, then, in the auditory SL task? Why is it that a task that is taken to reflect a domain-general capacity for registering distributional properties, either through TP computations (e.g., Endress & Langus, 2017; Endress & Mehler, 2009), or through chunk extraction (e.g., Perruchet, Poulin-Charronnat, Tillmann, & Peereman, 2014; Perruchet & Vinter, 1998), shows such peculiar patterns of modality, language, and stimulus specificity? The aim of the present study is to offer some novel insights regarding this important question.

1.1. The tabula rasa assumption

SL research often assumes the learner to be a tabula rasa, thereby viewing learning as the process of assimilating novel regularities. Following this assumption, the learning outcomes of an experiment are typically understood by considering the input structure alone. For example, if participants are presented during familiarization with an input containing 6 "words", with TPs of 1.0 between elements within words, their relative success in 2-AFC trials during the subsequent test phase is discussed by considering (1) the number of words in the stream, (2) the extent of the TPs between elements, and (3) the difference in TPs between "words" and foils in the test phase. The tabula rasa assumption is that the "words" (as well as the foils) were unknown to the participants at the start, so whatever is acquired (or not) during the familiarization session reflects the net efficiency of SL computations.

The tabula rasa assumption may indeed be true in many experimental designs when there is *no prior knowledge regarding co-occurrences of elements in the stream* (e.g., when learning abstract shapes, e.g., Turk-Browne et al., 2005; fractal visual stimuli, Schapiro, Gregory, & Landau, 2014, or novel cartoon figures, Arciuli & Simpson, 2011). However, in the domain of language, the tabula rasa assumption is unlikely. Humans hear speech from birth and start accumulating knowledge about the statistical properties of speech sounds in their native language by the hour. Here we claim that when participants perform an auditory SL task that utilizes verbal material, their existing representations regarding probabilistic co-occurrences of speech sounds in their native language impacts their performance on the task to a large extent. In a nutshell, we argue that one cannot predict the learning outcomes of an auditory SL task that contains linguistic elements, without weighing how the statistical properties of the input steam interact with participants' established expectations regarding the co-occurrences of speech sounds in their native language.

The suggestion that prior linguistic knowledge can modulate performance on auditory SL tasks is not entirely novel: It was raised as a possible explanation when accounting for discrepant results in the auditory SL task (and see Christiansen, Conway, & Curtin, 2000; Christiansen & Curtin, 1999, for an earlier version of this criticism). For example, whereas Perruchet and Poulin-Charronnat (2012) suggested that some peripheral factors of intelligibility of the speech stream could account for Endress and Mehler (2009) reporting no preference for words over phantom words in Italian speakers, Endress and Langus (2017) have raised the possibility that perhaps participants' prior experience in their native language (Italian vs. French) led to the discrepant findings (Footnote 3, p. 41). This issue, however, has critical importance, and cannot be left as a possible post hoc and open explanation for discrepant findings between laboratories. For if Endress and Langus (2017) are right, then the outcome of any study involving the learning of syllables during an auditory SL task, will be contingent on the sampled population. In other words, performance in the task does not simply reflect efficiency of SL computations as it was originally assumed, but reflects patterns of entrenchment of participants in their already established statistics.

The present paper focuses on this possibility by examining whether performance in the auditory SL task may be influenced by entrenchment. We define entrenchment as the influence of previously assimilated knowledge on the learning of the statistical properties from a new input. We examine this hypothesis by monitoring performance in SL tasks that implicate (or not) prior knowledge about the co-occurrences of patterns in the sensory stream. To preview our results, we show that the classical auditory SL task displays clear patterns of entrenchment. In contrast, SL tasks that do not involve prior knowledge regarding co-occurrence of elements are shown to be free of such entrenchment.

The hypothesis that SL performance is affected by entrenchment is compatible with two lines of existing work. First, there is a relatively large set of studies showing that the expectations that participants bring to SL tasks can be easily manipulated, affecting task performance. For example, pre-exposing participants to isolated words or part-words before the beginning of the familiarization stream has a dramatic effect on SL performance, which can either facilitate (Cunillera, Laine, Camara, & Rodriguez-Fornells, 2010; Lew-Williams, Pelucchi, & Saffran, 2011), or hinder (Perruchet et al., 2014; Poulin-Charronnat,

[1] Note that throughout the paper, unless noted otherwise, by auditory SL tasks we refer to tasks using auditory *verbal* material (e.g., Saffran et al., 1996), and by visual SL tasks we refer to tasks using visual non-verbal material (e.g., Kirkham et al., 2002).

[2] We refer here to the results of Experiment 2 from Erickson et al. (2016). In Experiment 1, zero correlations between different auditory SL tasks were also found, but these may be due to a small number of trials in each task, resulting in high measurement error (see Erickson et al., 2016, for discussion; see also Siegelman, Bogaerts, & Frost, 2016).

Perruchet, Tillmann, & Peereman, 2016) learning. In the same vein, pre-familiarizing participants with words of different length affects the size of the units they extract from the input (Lew-Williams & Saffran, 2012). Relatedly, studies that examined the learning of two consecutive sub-streams with different statistical properties, showed that learning one set of regularities affected subsequent learning (e.g., Gebhart, Aslin, & Newport, 2009; Karuza et al., 2016), and that this depends on the overlap between the statistical properties of the two stream (Siegelman, Bogaerts, Kronenfeld, & Frost, submitted). While none of these studies focused directly on the statistics that originate from participants' native language, they do show how SL performance is potentially affected by prior knowledge. If SL performance is so easily impacted by presenting participants with various statistics during the experimental session, exposure to language prior to the experiment (long-lasting exposure in the case of adults), should impact participants' performance to even a larger extent. A more direct source of support for the entrenchment hypothesis comes from studies suggesting that phonotactic cues characteristic of a language drive segmentation of the speech input. For example, Finn and Hudson Kam (2008) showed that, when the 'words' in the auditory stream presented to native English participants included illegal consonant sequences in English, segmentation did not concur with the TPs in the stream (and see Mersad & Nazzi, 2011; Onnis, Monaghan, Richmond, & Chater, 2005, for similar conclusions).

Here we drive these claims further. The entrenchment hypothesis suggests that prior knowledge impacts auditory SL performance in *any* experimental setting, not only when the stimuli chosen for the task directly clash with specific knowledge of one's native language. We thus argue that prior knowledge in any given language *always* raises predictions regarding probable co-occurrences of speech elements, and this influences performance in the auditory SL task, regardless of "words" selected for the experiment. To be clear, our claim is that performance in an auditory SL task may not reflect segmentation abilities exclusively, as is typically assumed, but may also reflect individuals' entrenchment in the statistics of their language gained through ongoing exposure to speech. This hypothesis offers a unified explanation for the list of puzzles we have outlined above. It would explain why performance in auditory and visual SL tasks is uncorrelated, explain why performance with one set of "words" in a familiarization stream does not necessarily predict performance with another set of words, it would explain why different developmental trajectories have been reported for auditory and visual SL, and it would explain why the same experimental design employed in different languages may result in different outcomes. The critical question, however, is how can our claim be empirically established?

1.2. Symptoms of entrenchment

Although it is possible to generate hypotheses regarding how the statistical properties of a native language result in predictions impacting continuous speech segmentation, a full theory of entrenchment requires investigations well beyond the scope of any single study. Such theory would not just center of TPs of syllables in a language, but should map all cues that could, in principle, impact speech segmentation, provide empirical evidence regarding the relative weights of each of these cues, and their possible interactions with one another. Then, through comprehensive corpora analyses, it would have to quantify the prevalence of these cues in the language, and finally, put these ranges of hypotheses to the test. To exemplify the deep complexity of this question, even if an accurate corpora analysis would produce a distribution of all TPs between syllabic segments in the language, there are other cues that could affect segmentation, such as the TPs of phonemic segments (e.g., Adriaans & Kager, 2010), higher order TPs between syllables (e.g., probability of C given both A and B; e.g., Thompson & Newport, 2007), backward TPs (e.g., Perruchet & Desaulty, 2008), or non-adjacent dependencies (e.g., Gómez, 2002; Newport & Aslin, 2004). Moreover, simple frequency of elements (phonemes, syllables, or larger chunks) should come into play as well (e.g., Thiessen, Kronstein, & Hufnagle, 2013), and then there are all the possible interactions between these cues.

A possible strategy to test the entrenchment hypothesis in SL, therefore, is to identify a possible symptom of entrenchment – an operational measure that can distinguish between situations where entrenchment does and does not play a role. This is the strategy we adopted here.

1.3. Internal consistency

When there is no prior knowledge whatsoever, and thus no possible predictions regarding the co-occurrence of elements in the stream, then all patterns are equal in terms of what they impose on the learner. Consider for example, an input stream with K patterns. If the patterns do not differ in terms of a priori predictions, then correlations of performance between these items should be high. This is labeled "internal consistency" – a situation in which all test items tap into the same construct. In contrast, if items do differ in terms of a priori knowledge, then the patterns in the stream will not be equal in terms of what they impose on the learner, and consequently some variance between patterns would emerge. The symptom of this state of affairs is a lower correlation in performance between items. In other words, with high internal consistency, learning Pattern A predicts learning Pattern B, whereas with low internal consistency, learning Pattern A would not necessarily predict learning Pattern B.

Operationally, the standard way to quantify internal consistency in a test is through the measure of Cronbach's α (Cronbach, 1951). According to test theory, Cronbach's α is *an estimate for the amount of shared variance across items*. As shown in the formula below, Cronbach's α is a function of the numbers of items in the test (K), their mean variance ($\bar{v}$), and the average covariance between them ($\bar{c}$).

$$\alpha = \frac{K\bar{c}}{(\bar{v} + (K-1)\bar{c})}$$

A critical clarification is required here: Cronbach's α is sensitive to whether items in the test tap the same theoretical construct, but is *not* affected by a simple manipulation of item difficulty. If two items measure the same theoretical construct (for example, TPs computation), but one item is more difficult in terms of computation (for example, by having a lower TP in the familiarization stream), the two items should still be highly correlated. This is because all participants who answered the more difficult item correctly, will also answer the less difficult one correctly. In contrast, if the items measure different constructs (for example, one mostly tapping TP computation, but another mostly affected by entrenchment in the statistics of the native language), success in one will not necessarily predict success in the other, and the variance in the test will be traced to two different sources. Hence, low internal consistency does not necessarily imply that something is wrong or unreliable with a given task, it simply shows that items in the task tap different abilities.

Our entrenchment hypothesis has very clear testable predictions. First, the visual SL task that uses novel abstract shapes does not implicate a priori predictions regarding co-occurrence of elements, and should therefore show high internal consistency. By contrast, if auditory SL performance implicates prior knowledge as we hypothesize, then this will be revealed by a lower internal consistency in the task, independent of overall performance in the task. Thus, in the auditory SL task, performance with one "word" will not necessarily predict performance with another "word". Second, the entrenchment hypothesis predicts that an auditory SL task that does not implicate prior knowledge regarding co-occurrence of elements will resemble the internal consistency of visual SL, but not the verbal auditory SL task. Third, the entrenchment hypothesis suggests that the zero correlation between auditory and visual SL performance (Siegelman & Frost, 2015), may not be due to modality constraints as was previously suggested (Frost et al.,

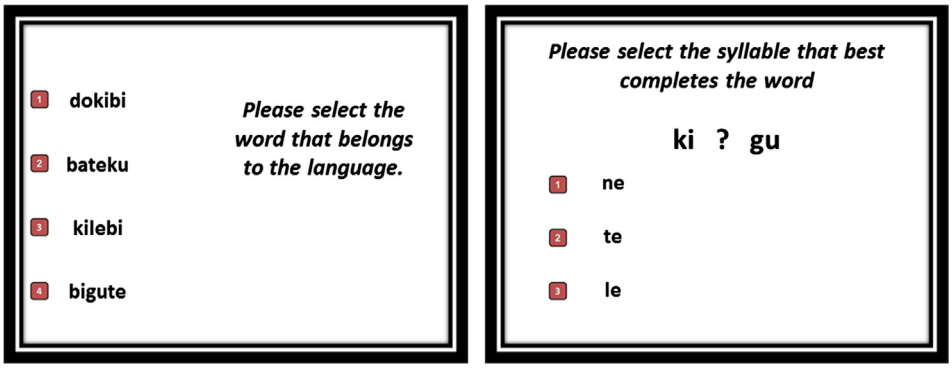


Fig. 1. Examples for test trials in Experiment 1a: A 4-AFC recognition trial (left), and a pattern completion trial (right). In all trials, stimuli were auditorily presented, one after the other, and their written forms appeared simultaneously.

2015), but to the difference sources of variance that come into play in the two tasks, one that involve entrenchment in prior knowledge, and one that does not. If this is the case, then performance in the visual SL task will be correlated with performance in an auditory task when neither task involves prior knowledge. The following series of experiments were set to test these predictions.

2. Experiment 1

Our initial prediction regarding high internal consistency in the visual SL task can be easily verified by considering the Cronbach α value that this task has produced. Recently, a visual SL task which employs abstract novel shapes was shown to withstand psychometric scrutiny by increasing the number of trials in the test, and expanding the range of difficulty of test items (Siegelman et al., 2016). The improved visual SL task was tested by Siegelman et al. (2016) in a sample of 62 participants. As hypothesized, the visual SL task produced a high Cronbach α value of 0.88. This represents a high score in line with typical psychometric standards (high internal consistency results in Cronbach α *values* around 0.8, e.g., Streiner, 2003), demonstrating that all items in the task equally tap the same construct – extraction of statistical properties.

The aim of Experiment 1 was to extend our investigation to two additional learning conditions (labeled Experiments 1a, 1b), with identical designs to the new visual SL task (Siegelman et al., 2016), but using different materials, to compare their internal consistency to that of learning abstract shapes.

First, in Experiment 1a, we employed an auditory verbal stream akin to the typical auditory SL task (Saffran et al., 1996). Our entrenchment hypothesis predicts that in contrast to learning novel shapes, low internal consistency would be revealed for this stream, due to participants' entrenchment in the statistics of co-occurrence of spoken segments in their language. Experiment 1b takes this strategy one step further. For this experiment, we generated auditory stimuli that do not implicate prior knowledge regarding co-occurrence of elements. We selected for this experiment familiar sounds as basic elements in the stream (e.g., glass breaking, dog barking, clock ticking, etc.). While participants are probably acquainted with each individual element, they likely do not have prior expectations regarding their co-occurrences. Our entrenchment hypothesis has then clear predictions: Although this will be an auditory task, paralleling the typical verbal auditory SL task, high internal consistency will emerge in this experiment, similar to the visual SL task.

2.1. Experiment 1a

2.1.1. Methods

2.1.1.1. Participants. Fifty-five students of the Hebrew University (22 males) participated in the study for payment or course credit. Participants had a mean age of 24.7 (range: 19–32), were all native speakers of Hebrew, and had no reported history of learning or reading disabilities, ADD or ADHD.

2.1.1.2. Materials, design, and procedure. The language included 16 CV syllables, which were synthesized in isolation using PRAT software (Boersma, 2001), at a fundamental frequency of 76 Hz and a syllable duration of 250–350 ms. Syllables were organized into 8 "words": 4 words with TPs = 1 (*munatu, bateku, modane, lodogi*) and 4 words with TPs = 0.33 (*kilegu, lekibi, biguki, gubile*). The 8 words were randomized to create a three-minute familiarization stream, which contained 24 repetitions of each word, without breaks between words (identical for all subjects). The only constraint in the randomization order was that the same word could not be repeated twice in a row. Prior to familiarization, participants were instructed that they would hear a monologue in an unfamiliar language, and that they would later be tested on their knowledge of the language. The monologue was then played to participants via earphones.

Following familiarization, a 42-item test phase began, identical in its design to the test described in detail in Siegelman et al. (2016). The first 34 trials were forced-choice questions, 22 trials with two options (2-AFC trials), and 12 trials with four (4-AFC). Trials included different foils varying in their level of difficulty (TPs of targets 0.33 or 1, foils with TPs ranging from 0 to 0.33), and tested knowledge not only of the full word-triplets (e.g., *biguki*), but also on pairs of syllables (e.g., *bigu* or *guki*). The 34 items in the forced-choice block were presented in a random order for each participant, and with a random order of the options within a trial (i.e., target and foil/s). In each trial, all options were played auditorily to participants, one after the other. Simultaneous to the auditory presentation, the written forms of each option were presented next to a number from 1 to 4 (see Fig. 1, left panel, for an example). Participants were instructed to choose the number next to the word which they think belong to the language. After the forced-choice trials, a block of 8 completion trials started. In each completion trial, a target pair or triplet was played (with its visual written form presented on the screen; see Fig. 1, right panel, for an example), but with one of the syllables replaced by white noise. Three options were then played one after the other (with their written forms appearing simultaneously), and participants were asked to choose the option that best completed the missing pattern. Overall test score in the task ranged from 0 to 42, based on the number of correct test trials. For the full details regarding the construction of foils and test trials, see Table 3 in Siegelman et al. (2016).

2.1.2. Results

The distribution of test scores in the auditory SL task is shown in Fig. 2. On average, participants answered correctly on 22.38 of 42 test trials (SD = 4.01). According to the binomial distribution (aggregating the different probabilities of correct responses for the different test-

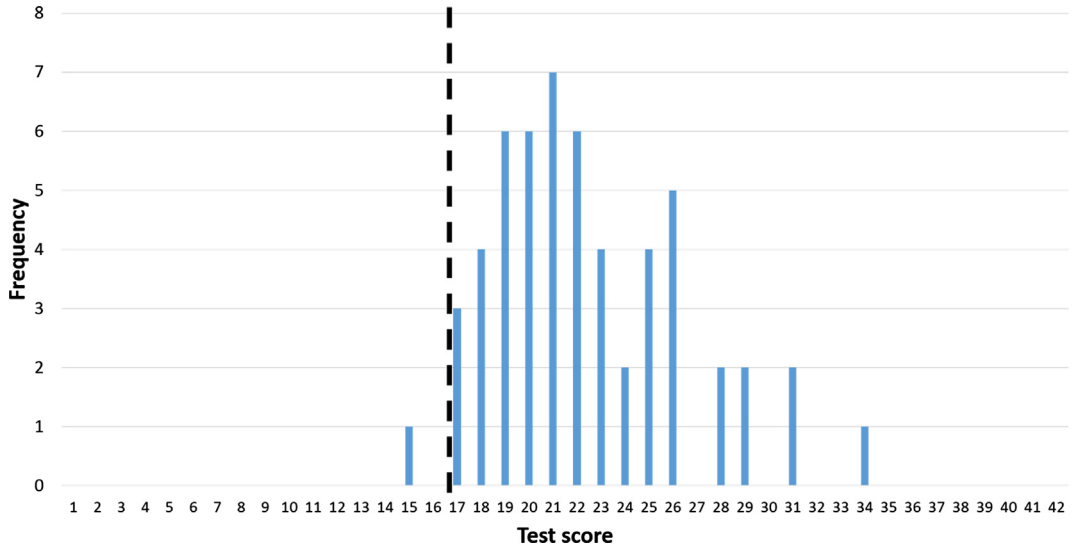


Fig. 2. Distribution of test scores in Experiment 1a (verbal auditory SL task). The dashed line represents chance-level performance (success in 16.67 trials).

items, i.e., aggregating across 2-AFC, 4-AFC and 3-AFC pattern completion trials), chance level performance in the task is 16.67 correct trials. One sample t-tests revealed a significant group-level learning in the task ($t(54) = 10.54$, $p < 0.001$).

2.1.2.1. Internal consistency. We next examined the internal consistency of the auditory SL task, estimating its Cronbach's α. This was done using the *alpha* function in *psych* package in R (Revelle, 2016), which calculates point estimates and confidence intervals for Cronbach's α, and using the *cocron* package (Diedenhofen & Musch, 2016), which performs significance tests for the comparison of Cronbach's α values across samples. As predicted from the entrenchment hypothesis, we found a very low estimate of $\alpha = 0.42$ (95% CI: [0.2, 0.64]) for the auditory SL task. This value fell well short of psychometric standards for task evaluation ($\alpha = \sim 0.8$, e.g., Streiner, 2003). Most importantly, this value presents significantly lower internal consistency compared to the Cronbach's α in the visual SL task from Siegelman et al., $\alpha = 0.88$ (95% CI: [0.83, 0.93]; comparison to the auditory SL: χ^2 (1) = 31.29, $p < 0.001$). To ascertain that this difference in internal consistency was not due in any way to the better performance in the visual SL task (26.4/42 trials correct vs. 22.38/42 trials correct in the auditory SL task, $t(1\ 1\ 5) = 3.25$, $p = 0.002$), we matched performance in the two tasks by removing the 12 best subjects in the visual SL, remaining with a sample of $n = 50$ with a mean performance of 23.4/42, no longer differing from performance in the auditory SL task ($t(1\ 0\ 3) = 1.01$, $p = 0.3$). The internal consistency of this sub-sample was indeed somewhat lower, $\alpha = 0.76$ (95% CI: [0.65, 0.86]). However, the difference in internal consistency between the auditory and visual SL tasks remained highly significant (χ^2 (1) = 9.07, $p = 0.003$).

In Experiment 1b we proceeded to examine the internal consistency of another similarly designed SL task, this time with non-verbal auditory sounds.

2.2. Experiment 1b

2.2.1. Methods

2.2.1.1. Participants. An additional sample of 62 students (20 males, mean age = 23.18, range: 19–34) at the Hebrew University was recruited for Experiment 1b. Similarly to Experiment 1a, all participants were native speakers of Hebrew, without a reported history of learning or reading disabilities, ADD or ADHD.

2.2.1.2. Materials, design, and procedure. The task had a similar design to that from Siegelman et al. (2016) and the verbal auditory SL in Experiment 1a. The only major difference was the materials used—this time, we selected 16 everyday familiar sounds from online repositories (http://www.bigsoundbank.com/, https://freesound.org/). All sounds were then manipulated using Audacity software to have a length of 800 ms. The 16 sounds are available online at: http://osf.io/x25tu.

Familiarization was identical to that in Siegelman et al. (2016). For each participant, the 16 sounds were randomly assigned to 8 triplets (4 with TPs = 1 and 4 with TPs = 0.33). Triplets were then randomized into a familiarization stream with 24 repetitions of each triplet, without immediate repetitions, with breaks of 200 ms between sounds both between and within triplets. Participants were instructed to listen carefully to the stream of sounds, as they would later be tested. The test phase was identical in its design to that of the visual SL and verbal auditory SL tasks, with 42 trials (34 forced-choice followed by 8 pattern completion trials). In each trial, options were played (auditorily) one after the other, with visual cues appearing on the screen next to the numbers signaling the corresponding keys on the keyboard (see Fig. 3 for examples). Possible scores again ranged from 0 to 42, based on the number of correctly identified targets.

2.2.2. Results

The distribution of test scores is shown on Fig. 4. Average performance was 23.5 trials correct out of 42 trials (SD = 5.6), which was significantly better than the task's chance-level of 16.67 ($t(61) = 9.59$, $p < 0.001$). Mean performance did not differ from the success in the verbal auditory SL task in Experiment 1a ($t(1\ 1\ 5) = 1.27$, $p = 0.21$).

Most importantly, and in line with our predictions, we found a high internal consistency for the auditory non-verbal SL task, with a Cronbach's α of 0.73 (95% CI: [0.6, 0.84]). This value was significantly higher compared to the verbal auditory SL task from Experiment 1a (χ^2 (1) = 7.89, $p = 0.005$). Moreover, it was almost identical to the internal consistency results with the visual SL task reported by Siegelman et al. (2016), when samples are matched in performance as in Exp. 1a (χ^2 (1) = 0.18, $p = 0.67$).

2.2.3. Discussion

Taken together, the results of Experiment 1a and 1b provide support for the entrenchment hypothesis. Both experiments involved auditory SL, with similar designs. However, their outcome in terms of internal consistency was dissimilar. Whereas the stream of syllables in Experiment 1a resulted in a very low value of internal consistency,

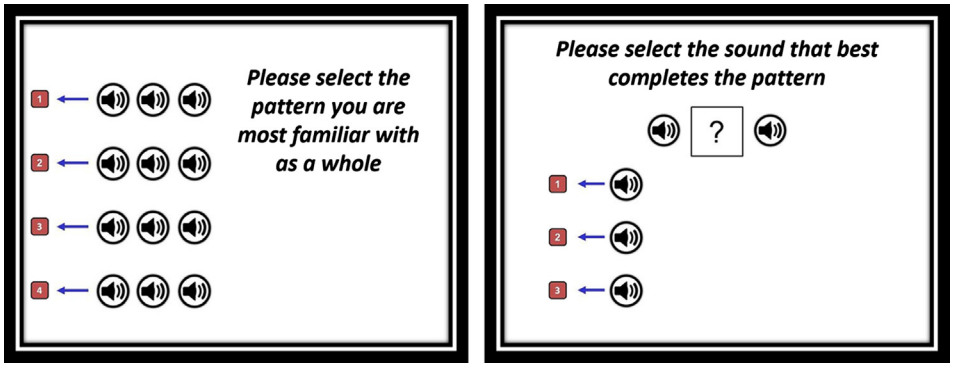


Fig. 3. Examples for test trials in Experiment 1b: A 4-AFC recognition trial (left), and a pattern completion trial (right). In all trials, stimuli were auditorily played to participants one after the other, and visual cues (speaker icons) appeared simultaneously.

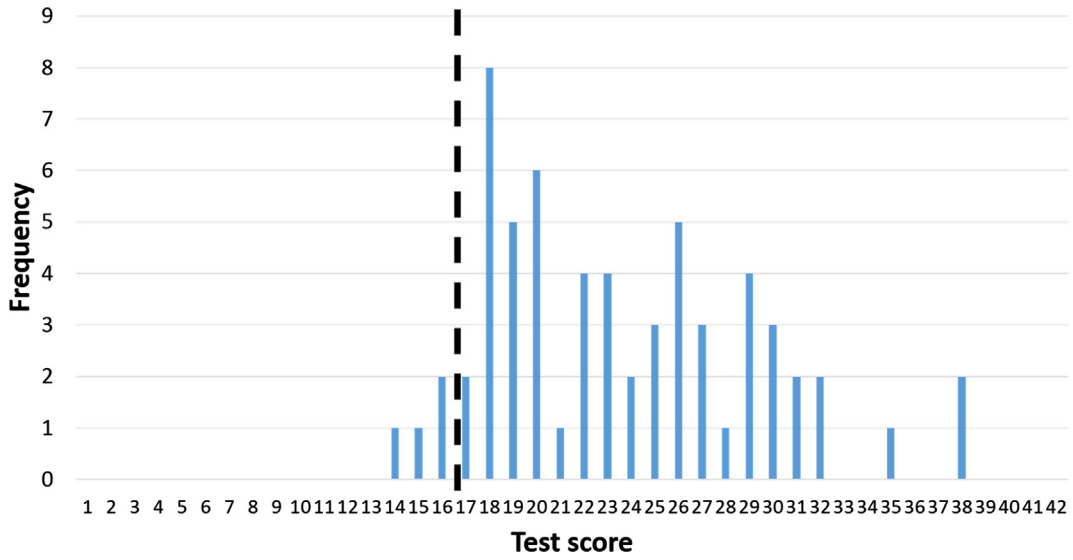


Fig. 4. Distribution of test scores in Experiment 1b (auditory non-verbal SL task). The dashed line represents chance-level performance (success in 16.67 trials).

simply substituting the syllables by non-verbal stimuli (Experiment 1b), led to high internal consistency. We emphasize that the critical difference between the two streams was the prior knowledge about the co-occurrences of the individual elements: a priori knowledge for verbal stimuli, no knowledge for the co-occurrence of non-verbal sounds. Note that this difference had no impact on the *overall* success in the tasks, which resulted in a similar level of performance. From a theoretical perspective, the effect of prior knowledge is not uniform across all items in the verbal auditory SL task. It could facilitate performance for some items but hinder performance for others, resulting in low internal consistency without necessarily impacting overall success (see also Poulin-Charronnat et al., 2016). Thus, the only difference between the verbal and non-verbal tasks was in the amount of shared variance between items, or, in the extent to which performance in one item predicted performance in other items.

One might wonder whether another possible factor - the length of familiarization - might have contributed to the difference in internal consistency between the tasks with verbal stimuli (i.e., auditory verbal SL) and those with non-verbal stimuli (i.e., visual SL, and the auditory non-verbal SL). While familiarization lasted 9.5 min in the non-verbal tasks, familiarization in the auditory SL task was shorter, around 3 min, because the individual syllables were shorter than the non-verbal material. We tested this hypothesis in a follow-up study (with a new sample of $n = 55$), with a similar task to that of Experiment 1a, but tripled familiarization length (72 repetitions of each word, 9 min overall). Still, internal consistency was very low (and numerically even *lower*): $\alpha = 0.27$.[3] This suggests that our results cannot be explained by familiarization length.

Considering the impact of modality, it seems that the internal consistency of the auditory non-verbal SL task more closely resembles that of the visual SL task with abstract shapes, rather than the verbal auditory SL. This would suggest that correlations in performance (or the lack of) are driven not by modality constraints (Frost et al., 2015), but by prior knowledge regarding co-occurrences of elements. We tested this hypothesis directly in Experiment 2.

3. Experiment 2

In a recent model explaining modality specificity effects in SL, Frost et al. (2015) have argued that the lack of correlation in performance in

[3] Note that in this follow-up we replaced two syllables from Experiment 1a (*ki* was changed to *ko*, *mo* was changed to *mu*). This was done given concerns that specific part-words in the original stream might resemble Hebrew words, potentially reducing the internal consistency of the task. This change had no effect on internal consistency, diminishing our concerns.

visual and auditory SL tasks stems from different constraints in processing regularities in the visual and auditory cortices. The entrenchment hypothesis offers an alternative explanation for this lack of correlation. This, again, sets clear predictions. If the zero correlation between visual SL and auditory verbal SL stems from differences in prior knowledge regarding element co-occurrence, then individual performance in the non-verbal visual SL task should correlate with individual performance in the auditory non-verbal task. We tested this prediction in Experiment 2.

For this experiment, we re-tested the participants of Experiment 1b on the non-verbal auditory task, and more importantly, tested them with the visual SL task (Siegelman et al., 2016). This provided us first, with a measure of stability of performance in the auditory non-verbal SL task, and second, with a measure of shared variance in performance in two SL tasks that implicate different modalities, but do not implicate prior knowledge.

3.1. Methods

All subjects of Experiment 1b were re-contacted and invited to return to the lab for a follow-up study in return for course credit or payment. Forty-two participants (11 males; mean age 22.76, range: 20–28) replied positively. In this session, participants were first re-tested on the auditory non-verbal task from Experiment 1b, and then undertook the visual SL task from Siegelman et al. (2016). Note that for the auditory task, while the sounds used in Experiment 2 were the same as those in Experiment 1b, the triplets during familiarization were re-randomized for each participant. The mean interval between the initial testing session (Experiment 1b) and retest (Experiment 2) was 93.7 days (SD = 28.18, range: 54–158).

3.2. Results

3.2.1. Test-retest

Mean performance on the re-test of the auditory non-verbal SL task was 20.73/42 (SD = 6.2), which was significantly better than chance ($t(41) = 3.81$, $p < 0.001$). Fig. 5 shows the test-retest scatter plot of scores in the two sessions. Test-retest reliability was high, estimated at 0.7 (95% CI: [0.5, 0.83]), a value similar to the reported test-retest reliability of the visual SL by Siegelman et al. (0.68, 95% CI: [0.48, 0.81]). This shows that performance in the auditory non-verbal task provides a stable signature of SL individual-level performance, and hence can be used to accurately estimate correlations with other measures (see Siegelman et al., 2016, for a detailed discussion). It is worth noting that, surprisingly, mean performance at re-test was for some reason *lower* than the performance of the same sub-sample on the first administration (20.73 vs. 23.73, $t(41) = 3.81$, $p < 0.001$). This, however, is peripheral to our investigation since such interference should, if anything, lead to an *under*estimation of the observed correlation with visual SL. It is also worth noting that high internal consistency was again observed for the auditory non-verbal task, with $\alpha = 0.76$, replicating the finding from Experiment 1b.

3.2.2. Visual-auditory correlations

The mean success rate in the visual SL was 26.04/42 trials correct (SD = 8.4), similar to that reported in Siegelman et al. (2016) of 26.4/42 ($t(1\,0\,2) = 0.19$, $p = 0.85$). Importantly, the main research question of this experiment was whether a correlation in performance would be found across modalities. Fig. 6 presents the correlation between visual SL and the auditory non-verbal SL task scores. As can be seen, and in line with our entrenchment hypothesis, a significant correlation between the tasks was revealed, of $r = 0.55$ (95% CI: [0.3, 0.73]). A similar correlation was found between the visual SL and the scores of the auditory non-verbal SL task in the first administration ($r = 0.5$ (95% CI: [0.23, 0.7]). Together, the strong, positive correlation of SL performance across modalities stands in contrast to the findings by Siegelman and Frost (2015), reporting a zero correlation between visual SL and verbal auditory SL.[4]

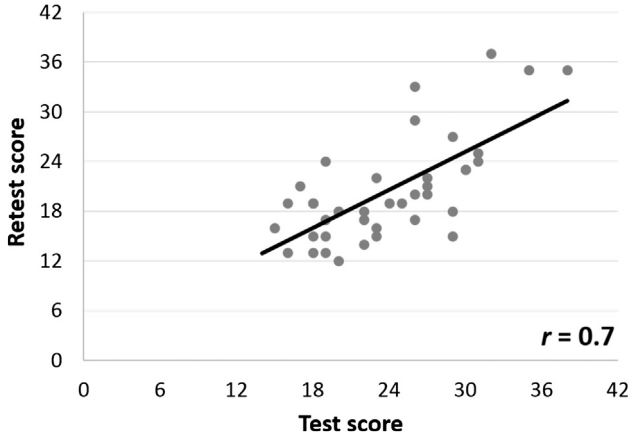


Fig. 5. Test-retest reliability of the auditory non-verbal SL task.

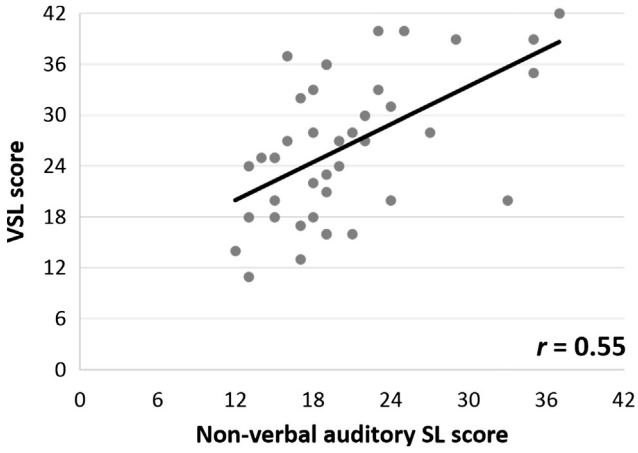


Fig. 6. Correlation between the auditory non-verbal SL task and the visual SL task.

4. Experiment 3

The aim of Experiment 3 was twofold. First, given the theoretical importance of our main claims, we wanted to ensure that the previous observed differences in internal consistency between the verbal auditory SL and visual SL were not due to idiosyncratic properties of the task developed by Siegelman et al. (2016) or the "words" employed in the verbal auditory SL task (e.g., their specific syllabic structure, or their acoustic properties). We therefore sought to replicate the dissociation in internal consistency between the visual SL and the verbal auditory SL tasks, with different sets of stimuli, using a more standard variant of these tasks (based on Saffran, Newport, Aslin, Tunick, & Barrueco, 1997, for auditory SL; Turk-Browne et al., 2005, for visual SL). Hence, in both visual and auditory SL we employed two new stimuli conditions, each with 6 triplets (all with TPs of 1.0) and with a test consisting of 36 2-AFC trials comparing triplets to foils with TPs of 0. Our entrenchment hypothesis predicts that entrenchment would impact

[4] An interesting related question is, then, how much variance exactly is shared between the non-verbal tasks in the two modalities. Note that the observed correlation of $r = 0.5$, does not take into account the imperfect reliability of the two tasks. More formally, the correlation between two variables is upper-bound by the square root of the product of their reliability ($\rho_{xy} \leq \sqrt{\rho_{xx} * \rho_{yy}}$). When taking into account the measures' reliability, using Spearman's correction for attenuation formula, $\rho_{x'y'} = \frac{\rho_{xy}}{\sqrt{\rho_{xx} * \rho_{yy}}}$, the correlation of 0.55 points to an expected correlation of 0.79, hence 62% of shared variance.

internal consistency for any set of linguistic stimuli, hence in Experiment 3 we used two novel sets of syllables. The second goal of Experiment 3 was to employ triplets that were constant across all participants in all tasks. This had both a methodological and a theoretical motivation. Methodologically, we aimed to rule out the possibility that the difference in internal consistency between verbal (Experiment 1a) and non-verbal (visual SL, Siegelman et al., 2016, and the non-verbal auditory SL task, Experiment 1b) tasks was due to the different randomization procedure in the two tasks (fixed 'words' in the auditory verbal SL, but random triplets in the visual SL). From a theoretical perspective, employing fixed triplets across conditions enabled us to pinpoint, for the first time, how each triplet in the familiarization stream contributed to the variance in task performance across our sample of participants.

4.1. Methods

4.1.1. Participants

A sample of 200 Hebrew University students (68 males), who did not take part in Experiments 1 or 2, participated in this study. They had a mean age of 23.68 (range: 19–31). Similarly to Experiments 1 and 2, participants were all native speakers of Hebrew, and declared no history of learning or reading disabilities, ADD or ADHD. Participants were assigned to participate in either the visual or auditory SL task ($n = 100$ in each), and then within each modality, they were assigned to one of two stimuli conditions ($n = 50$ in each of the stimuli conditions of the auditory SL; $n = 51$ and $n = 49$ in stimuli condition 1 and 2 of the visual SL, respectively. The number of participants was not fully identical in the two conditions of the visual SL due to an experimenter error).

4.1.2. Materials, design, and procedure

Auditory SL task. Both stimuli conditions of the auditory SL task had an identical design, but with different materials (i.e., different syllables and "words"). Each language consisted of 18 syllables. In stimuli condition 1, the material was generated akin to that from Experiment 1a: syllables that were synthesized in isolation using PRAT (Boersma, 2001), at a fundamental frequency of 76 Hz and a syllable duration of 250–350 ms. In contrast, stimuli in condition 2 were based on naturally-spoken syllables, which were recorded by a native speaker of Hebrew. Importantly, syllables were recorded in isolation, to avoid any prosodic cues for segmentation. The syllables were 220–360 ms long, and ranging in frequency between 140 Hz and 190 Hz.

In each stimuli condition, the 18 syllables were then organized into 6 words (constant across all participants), all with within-word TPs of 1 (see Table 1). The 6 words were then randomized to create a familiarization stream containing 24 repetitions of each word, without breaks between words (word order in familiarization was identical for all subjects in each condition), with the only constraint of no immediate repetitions. Familiarization instructions were similar to Experiment 1a: participants were told they would hear a monologue in an unfamiliar language, and that they would later be tested on their knowledge of the language. The test phase included 36 2-AFC trials, each containing a pair of stimuli: a "word", and a foil (always with TPs = 0; see Table 1).

Table 1
Words and foils in the two auditory SL stimuli conditions in Experiment 3.

Stimuli condition 1		Stimuli condition 2	
Triplets (TPs = 1)	*Foils (TPs = 0)*	*Triplets (TPs = 1)*	*Foils (TPs = 0)*
lenamo	lerifa	dukeva	dulize
mivofa	minade	kutoze	kugabe
nubogi	nukaro	nigobe	nitomu
paluro	pabose	nolita	nodiva
saride	savogi	sogamu	sokeba
tikase	tilumo	vudiba	vugota

The 36 test trials were presented in random order with a constraint that the same word or foil could not appear in two consecutive trials. In each trial, participants heard the two options (i.e., a word and a foil) one after the other in a random order (with an ISI of 1000 ms), and were asked to decide which tri-syllabic sequence belonged to the language by pressing 1 or 2 on the number pad to select either the first or the second word. Scores in the test ranged from 0 to 36, based on the number of correctly identified words over foils.

Visual SL task. The visual SL was similar in its design to the auditory SL but with visual-nonverbal, rather than auditory-verbal, material. Here also there were two stimuli conditions, one with 16 shapes (e.g., Fiser & Aslin, 2001; Turk-Browne et al., 2005), and one with 16 Ge'ez letters, which were unfamiliar to participants (e.g., Karuza, Farmer, Fine, Smith, & Jaeger, 2014). The 6 triplets in each stimuli condition of the visual SL are presented in Table 2. Similar to the parallel auditory SL condition, triplets were fixed across subjects. Familiarization again included 24 repetitions of each triplet (in an identical order across participants), without immediate repetitions of triplets. Exposure duration was 600 ms per shape, with an ISI of 100 ms (both within- and between- triplets). Participants were instructed to attend the familiarization stream, as they would later be tested. The test phase included 36 trials (presented in a random order), each comprising of a triplet and a foil (all foils with TPs = 0, see Table 2). In each trial the triplet and foil appeared one after the other in a random order (with a 1000 ms break between options), and participants were asked to choose which of the two options they are more familiar with (as a sequence).

4.2. Results

4.2.1. Mean performance and internal consistency

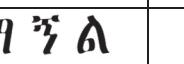
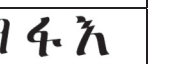
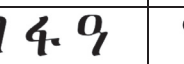
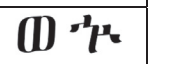
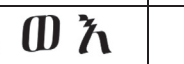
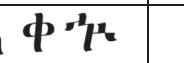
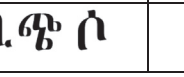
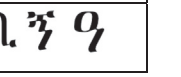
Performance in the auditory SL task was quite similar in the two conditions, 24.16/36 (67.1%) for stimuli condition 1, and 23.84/36 (66.2%) for stimuli condition 2. Both values represent group-level learning, significantly differing from the chance level of 50% (condition 1: $t(49) = 9.52$, $p < 0.001$; condition 2: $t(49) = 10.69$, $p < 0.001$). Mean performance rates were similar in the parallel visual SL tasks, with 23.86/36 (66.3%) correct trials for stimuli in condition 1, and 23.01/36 (63.9%) for stimuli in condition 2, again showing significant learning (condition 1: $t(50) = 6.45$, $p < 0.001$; condition 2: $t(48) = 6.01$, $p < 0.001$).

Our main focus, however, was the internal consistency values. As predicted by the entrenchment hypothesis, internal consistency was high in both visual conditions: $\alpha = 0.84$ (95% CI: [0.75, 0.91]) for abstract shapes, and $\alpha = 0.78$ (95% CI: [0.67, 0.87]) for Ge'ez letters. The internal consistency in the two auditory conditions was as hypothesized poorer, $\alpha = 0.54$ (95% CI: [0.36, 0.73]) for condition 1, and $\alpha = 0.59$ (95% CI: [0.43, 0.77]) for condition 2, albeit somewhat higher than that of Experiment 1a. Significance tests revealed a difference between condition 1 in the visual SL to both auditory SL conditions (comparison to auditory condition 1: χ^2 (1) = 12.44, $p < 0.001$; comparison to auditory condition 2: χ^2 (1) = 9.97, $p = 0.002$), and a similar difference between visual SL condition 2 and the two auditory SL conditions (comparison to auditory SL condition 1: χ^2 (1) = 6.06, $p = 0.01$; comparison to auditory SL condition 2: χ^2 (1) = 4.35, $p = 0.04$). There was no difference in internal consistency between the two stimuli conditions within each modality (visual: χ^2 (1) = 1.17, $p = 0.28$; auditory: χ^2 (1) = 0.15, $p = 0.7$). Together, these results replicate the observed pattern in Experiment 1, in a more common variant of SL tasks, using different materials.

4.2.2. Factor analysis

Next, we sought to trace the underlying components of variance in the SL tasks, using an exploratory factor analysis. As noted above, targets and foils were fixed across participants within each experimental condition, to allow us to examine whether trials with specific

Table 2
Triplets and foils in the two visual SL stimuli conditions in Experiment 3.

	Stimuli condition 1		Stimuli condition 2	
No.	*triplets (TPs=1)*	*foils (TPs=0)*	*triplets (TPs=1)*	*foils (TPs=0)*
1				
2				
3				
4				
5				
6				

targets (or foils) map into common underlying components. We had two main predictions. First, we predicted that the variance explained by the leading factor in the visual SL tasks would be larger than the variance explained by the main factor of the auditory SL task.[5] Second, we predicted that since all trials in the visual SL tasks tap the same component – the ability to extract transitional statistics from the input, all (or most) trials will correlate with the main factor. In contrast, in the auditory SL tasks, entrenchment will result in a non-uniform distribution of correlations. Trials related to some triplets will be loaded with the leading factor, whereas trials related to other triplets will not.

Appendices 1a and 1b present the full output of the factor analysis on the visual SL tasks, and Appendices 2a and 2b present the results of the factor analysis on the auditory SL tasks. The results of these analyses confirmed both of our predictions. First, the primary factor in the visual SL tasks accounted for 17.1% of the observed variance in condition 1, and for 14.6% in condition 2. In contrast, the primary factor in the auditory SL tasks accounted for 10.6% of the observed variance in Condition 1, and for 10.1% in Condition 2. Second, as hypothesized, in both conditions of the visual SL task virtually all trials, across different targets and foils, were positively loaded on the primary factor (35/36 trials in Condition1, 34/36 trials in Condition 2). In contrast, the auditory SL task presents a very mixed picture. In Condition 1, 14/36 trials were negatively loaded on the primary factor, and in Condition 2, 10/36 trials were negatively loaded on the primary factor. This points to different sources of variance explaining performance in the task.

Our factor analyses show how this methodology can be used to pinpoint traces of variance of different "words" in the stimuli set. For example, in condition 1 of the auditory SL task, *lenamo*, *mivofa, paluro* and *saride* had positive loadings on the leading factor (22 out 24 items related to these targets were positively correlated with it), while all 12 trials with the targets *nubogi* and *tikase* were *negatively* loaded on this same factor. This exemplifies that success in learning *nubogi* or *tikase*, not only does not predict success in learning *lenamo* or *mivofa*, but is in fact orthogonal to it. This indicates the main characteristic of entrenchment: not all patterns are alike when participants enter the learning situation. Admittedly, we do not have a clear account which characteristics of these words make them easier to perceive – that will require a detailed analysis of co-occurrence statistics of the linguistic environment of our speakers. However, as a first step, we examined a simpler prediction of the entrenchment hypothesis: that words that are learned better in verbal auditory SL tasks better resembles the prior linguistic knowledge of the learners. Experiment 4 was set to examine this prediction.

[5] Exploratory factor analyses by default produce more than one factor. Here we focus on how much of the variance is explained by the primary factor.

5. Experiment 4

Experiment 4 was set to further demonstrate the effect of prior knowledge on auditory verbal SL performance. This was done by examining an additional prediction of the entrenchment hypothesis, namely, that native speakers of the same language should show similar variation in SL accuracy outcomes, given the overlap between their existing knowledge from their native language and the stimuli used in the SL task. We tested this prediction by quantifying the resemblance of the verbal auditory SL stimuli to linguistic units in participants' native language. In order to do so, we recruited an independent sample of native Hebrew speakers, who ranked the stimuli from the previous verbal auditory SL experiments in this paper on their similarity to Hebrew. We predicted that these rankings would explain unique variance in the verbal auditory SL performance observed in the previous experiments in this paper. Specifically, we predicted that SL performance will be higher on "words" that are more Hebrew-like compared to "words" that do not resemble Hebrew. We also examined whether foils' resemblance to Hebrew would have a similar effect on SL performance.

5.1. Methods

5.1.1. Participants

Fifty students of the Hebrew University (14 males), who did not participate in any of the previous experiments, participated in this study for payment or course credit. Their mean age was 23.2 (range: 18–32), they were all native speakers of Hebrew, and had no reported history of learning or reading disabilities, ADD or ADHD.

5.1.2. Materials, design and procedure

All stimuli – both targets and foils - from verbal auditory SL tasks in Experiments 1a, follow-up of Experiment 1a (henceforth, 1a-FU), and Experiment 3 (condition 1) formed the materials for this experiment. Note that targets included both words (e.g., *bateku*) as well as part-words: pairs of syllables with high TP serving as targets (e.g., *bate*). This resulted in seventy-nine stimuli overall. All of these stimuli were comprised of syllables synthesized in isolation, with durations of 250–350 ms. The syllables used in Condition 2 of Experiment 3 were the recording of a human voice (native Hebrew speaker), and hence were

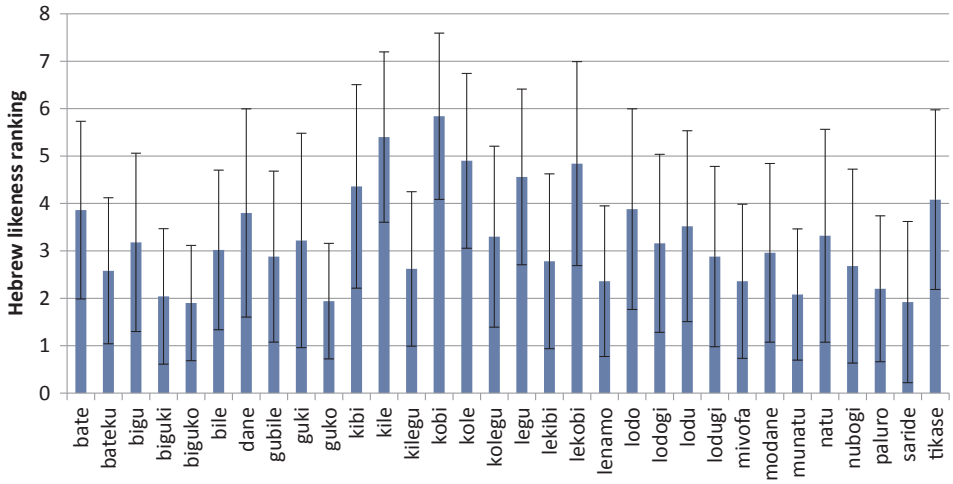


Fig. 7. Average rankings for auditory SL targets (error bars represent SD).

not included in order to maintain uniformity between stimuli.

An online ranking task was built using the Qualtrics platform, version 12/2017 (Qualtrics, Provo, UT). Participants did the task online from home. They were instructed to use earphones and sit in a quiet room when conducting the experiment. Before the beginning of the task, participants were told they would hear a robot speaking in a robot language, and that they need to rank each of the robot's words based on its similarity to Hebrew. A Likert-scale was used (1 for *not similar at all* and 7 for *very similar*). Participants were asked to try to use the entire range of the scale. In each trial, a single auditory stimulus was played automatically (participants could re-play the stimuli if they wished). Then, the participant ranked the stimulus by choosing one of the seven numbers and clicked "Next" to proceed to the next trial. After ranking the 79 stimuli, participants were asked to provide information regarding their gender, age, native language and the other languages they speak. The task took in total 5–10 min.

5.2. Results

5.2.1. Targets' and foils' rankings

Mean rankings for targets and foils are shown in Figs. 7 and 8 respectively. On average, the mean ranking of stimuli was 3.27, with substantial variance of 1.22 (range: 1.76–5.98). Note that rankings of targets and foils did not differ (targets' mean = 3.26, SD = 1.04, foils' mean = 3.27, SD = 1.15; $t = 0.968$, $p = 0.33$). Importantly, the presence of substantial variance in the rankings demonstrates that not all stimuli are experienced alike: some are experienced as very Hebrew-like while others are not.

5.2.2. Rankings as a predictor of auditory SL performance

To examine whether the similarity of the verbal auditory stimuli to Hebrew affected participants' SL performance, we used a logit mixed model including the targets' and foils' rankings as predictors of SL performance in forced-choice questions from Experiments 1a, 1a-FU, and 3. In 4-AFC trials, the average ranking of the three foils was used. The data thus included 34 trials for each subject from Experiments 1a and 1a-FU, and 36 trials for each subject from Experiment 3.

As the response in each trial was categorical (correct/incorrect), we used a logistic mixed-effect model, using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). The fixed effects in the model were standardized target ranking and foil ranking, as well as the following control variables: experiment (1a, 1a-FU or 3, dummy coded), question type (2-AFC or 4-AFC), target TP (0.33 or 1), and foil TP (range: 0–0.5). Note that target and foil rankings were standardized within each experiment separately (i.e., for each stimulus, we computed a standardized ranking score based on the mean rankings and SD in each experiment), due to acoustic differences in stimuli across experiments and the different context in which each stimulus was presented. The model also included a by-subject random intercept, which was the maximal random effect structure that converged (Barr, Levy, Scheepers, & Tily, 2013).

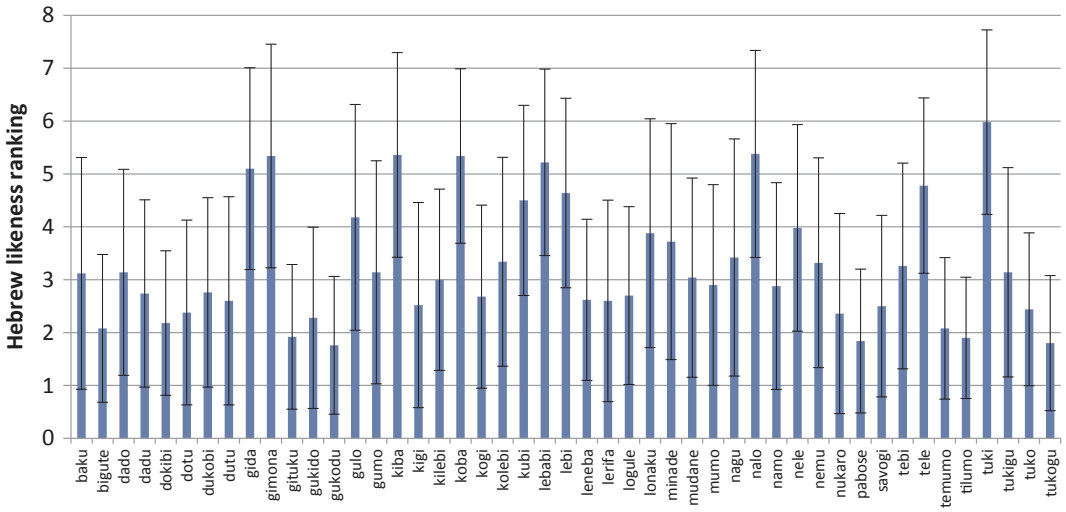


Fig. 8. Average rankings for auditory SL foils (error bars represent SD).

Table 3
Summary of the fixed effects in the mixed-effect logit model of Experiment 4.

Predictor	Coefficient (β)	SE	*z*	*p*
Intercept	−0.115	0.103	−1.123	.261
Target ranking	**0.102**	**0.029**	**3.54**	**< .001**
Foil ranking	**0.095**	**0.029**	**3.223**	**< .002**
Question type	0.871	0.07	12.316	< .001
Word TP[1]	−0.155	0.104	−1.495	.134
Foil TP	−2.163	0.318	−6.791	< .001
Experiment 2	0.04	0.083	0.491	.623
Experiment 3	0.129	0.1	1.291	.196

Note. Coefficients refer to a change of β in the logit probability of getting a correct response with every one-unit increase in the predictor. Bold rows highlight the effects of interest - Targets' and foils' resemblance to participants' native language.

[1] Surprisingly, we did not observe a significant effect of transitional probabilities of the words ($\beta = -0.11$, $z = -1.096$, $p = .272$). This stands in contrast to findings with non-verbal stimuli, in which word TP is a stronger predictor of performance (e.g., Bogaerts, Siegelman, & Frost, 2016; Siegelman et al., 2016). This, again, shows that performance on tasks with verbal material cannot be solely explained by the distributional properties of the input within the experimental session, but rather is affected by other factors – such as entrenchment.

The full output of the model is presented in Table 3. In line with our predictions, there was a significant effect for target ranking ($\beta = 0.102$, $z = 3.54$, $p < .001$) as well as for foil ranking ($\beta = 0.095$, $z = 3.223$, $p < .002$). This shows that participants performed better on trials including targets more similar to their native language, but also on trials with Hebrew-like foils. We interpret these results to suggest that when given a forced choice in a test phase, subjects were better able to select targets over foils when they were rated as more similar to Hebrew. In addition, subjects were better at eliminating foils that were more Hebrew-like, and determine they did not appear in the stream. Together, the results of Experiment 4 show that entrenchment is reflected not only in the correlation across items (e.g., internal consistency, Experiment 1–3), but also in auditory SL performance for different targets and foils.

6. General discussion

The original findings of Saffran et al. (1996), focused on how language is learned given the statistics of the input presented in the experimental session. Humans, however, learn the regularities of their language continuously from birth. Thus, when they come into the learning situation, even at an early age, and are presented with "novel words" (e.g., "bateku", "modane"), they are already entrenched in their language's statistics. These determine not only the learning outcomes, but also the learning process. The entrenchment hypothesis offers a unified account for some of the unsettled findings in SL research. It explains and, importantly, predicts (at least to some extent) when and why correlations in SL performance would be obtained, or not. It also explains why different outcomes have been reported across linguistic environments, samples, and materials.

The present set of five SL experiments was designed to examine how prior knowledge regarding co-occurrences of elements in continuous sensory streams would be reflected in the learning outcomes. In Experiments 1–3, we focused not on mean success rate, as most SL studies do, but rather on shared or distinct components of variance in performance, either within a task (i.e., internal consistency, factor analysis), or across tasks (i.e., between-task correlations). Our results are straightforward. We found that learning situations that do not involve prior knowledge regarding co-occurrence of elements are characterized by high internal consistency of learned items, regardless of modality. In contrast, when learning involves linguistic material, prior knowledge of participants leads to low internal consistency. Thus, success in recognizing "bidaku" in the stream does not necessarily predict success in recognizing "padoti", or "golabu". We also found that when learners are "tabula rasa" regarding co-occurrences of elements, significant correlation in SL performance is revealed even when two learning situations involve different modalities. Experiment 4 provides direct evidence for the entrenchment hypothesis, showing that variance in auditory verbal SL performance can be predicted by the resemblance of stimuli to participants' native language.

Given the theoretical implications of our findings, we sought to validate our claims by considering datasets from other laboratories that used in parallel a visual SL and auditory (verbal) SL tasks, with a similar design, for which internal consistency levels can be compared. We gained access to the full data of two such studies: Glicksohn and Cohen (2013), who had a sample of $n = 32$ adults in each task, and Raviv and Arnon (2017), who used a sample of $n = 125$ children (ages 6–12) in each task. Calculating the internal consistency in these two studies yielded the following results: In Glicksohn and Cohen (2013) the visual SL task had a Cronbach's α of 0.78 (95% CI: [0.65, 0.88]), while the auditory SL had a Cronbach's α of 0.39 (95% CI: [0.04, 0.66]). In Raviv and Arnon (2017), the visual SL had a Cronbach's α of 0.64 (95% CI: [0.54, 0.74])[6] compared to 0.25 (95% CI: [0.06, 0.44]) in the auditory SL. In both studies there was a significant difference in internal consistency between the visual and auditory SL (Glicksohn & Cohen, 2013: χ^2 (1) = 7.21, $p = 0.007$; Raviv & Arnon, 2017: χ^2 (1) = 15.03, $p < 0.001$). Thus, it seems that our findings regarding the internal consistency of visual SL versus auditory verbal SL indeed generalize to other experimental settings.

Taken together, the present study shows the critical effect of prior knowledge in determining SL outcomes. This has important implications for SL research. First, it sets a demarcation line between two types of learning situations, one when learning starts at zero, and one when it does not. The trajectory of learning may be quite different in these two settings. This also means that, methodologically, tasks that implicate prior knowledge such as the auditory verbal SL task cannot be easily borrowed to compare different samples of participants. Moreover, even within a sample of participants, comparing performance across learning conditions with different streams may sometimes be problematic. This is because the specific selection of "words" (and, possibly, foils), may manipulate not only the statistical information present in the stream, but also tap different expectations of participants given their entrenchment in prior statistics of their language (see the results of Erickson et al., 2016). Relatedly, from an individual-differences perspective, auditory SL tasks involving verbal material may not be the best proxy of net SL computations, because performance is also affected by participants' prior entrenchment regarding the specific stimuli in the task. This suggests that for predicting abilities related to SL (e.g., L2 leaning, syntactic processing, reading abilities), tasks should preferably not involve prior knowledge. Indeed, the non-verbal visual SL task has proven very useful in predicting individual differences in L2 learning (Frost, Siegelman, Narkiss, & Afek, 2013), knowledge of grammatical structure (Kidd & Arciuli, 2016), and reading ability (Arciuli & Simpson, 2012).

However, while setting the demarcation line between learning situations for which learning starts at zero and for which learning starts with prior knowledge regarding item co-occurrences, it is important to emphasize that organisms learn most regularities of their environment continuously. Therefore, SL in the real world involves in most cases the *updating* of prior statistics for upcoming predictions, rather than establishing entirely novel representations. This suggests that understanding SL from an ecological perspective, and specifically its role in language learning, requires advancing towards a mechanistic and

[6] It is worth mentioning that the visual SL task in Raviv and Arnon (2017) was based on a similar visual SL task by Arciuli and Simpson (2012), which also had a high internal consistency value of 0.79 in a sample of $n = 37$ adults.

detailed theory of entrenchment. In that sense, SL research should focus on providing systematic data regarding how prior expectations of a range of possible cues for learning are weighted together with the statistics of the input, to produce the learning outcomes of a given learning situation. Such research should not be limited only to co-occurrences of elements, but also to their interaction with a range of other linguistic cues such as prosody (Thiessen & Saffran, 2003), or phonotactics (e.g., Onnis et al., 2005). Such data can then be used to formulate a detailed computational model for the updating of existing representations during exposure to new input. One promising avenue can be the incorporation of Bayesian models, which weight prior expectations and new evidence equally, into SL research (see, e.g., Goldwater, Griffiths, & Johnson, 2009).

Finally, our data also shed light on recent debates regarding domain-generality vs. domain-specificity in SL (Conway & Christiansen, 2005; Frost et al., 2015; Milne, Petkov, & Wilson, 2017). The fact that a significant correlation was found in visual and auditory SL for material not involving prior knowledge, suggests that there are some common computations across modalities. This does not imply that one unitary device drives SL (cf. Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014; see also Arciuli, 2017). It does, however, open research avenues for investigating when and to what extent SL computations are similar across domains. To emphasize, such research should not only consider modalities, but also materials and the prior knowledge they implicate.

Acknowledgments

This paper was supported by the ERC Advanced grant awarded to Ram Frost (project 692502-L2STAT), the Israel Science Foundation (Grant 217/14 awarded to Ram Frost) and by the National Institute of Child Health and Human Development (RO1 HD 067364 awarded to Ken Pugh and Ram Frost, and PO1 HD 01994 awarded to Haskins Laboratories). Louisa Bogaerts received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 743528 (IF-EF). We wish to thank Inbal Arnon for comments and helpful discussion. We also thank Arit Glicksohn and Limor Raviv for kindly sharing their data.

Appendices

Appendix 1a. Exploratory factor analysis on the data from the visual SL task in Experiment 3, condition 1. Included here are loadings of all trials on the main three extracted factors. Numbers under 'target' and 'foil' correspond to the presented stimuli in Table 2 above.

Serial trial no.	Target word	Foil	Factor 1 loading	Factor 2 loading	Factor 3 loading
1	1	1	0.212	0.287	0.222
2	1	2	0.472	0.095	0.234
3	1	3	0.176	0.415	0.420
4	1	4	0.528	−0.003	0.119
5	1	5	−0.096	−0.357	0.451
6	1	6	0.684	0.073	−0.137
7	2	1	0.336	0.423	−0.275
8	2	2	0.639	0.119	−0.244
9	2	3	0.261	0.595	0.159
10	2	4	0.495	−0.268	−0.009
11	2	5	0.218	−0.242	0.165
12	2	6	0.580	−0.100	−0.207
13	3	1	0.503	−0.183	−0.002
14	3	2	0.437	−0.239	−0.077
15	3	3	0.302	0.587	0.134
16	3	4	0.303	−0.419	0.060
17	3	5	0.451	−0.323	0.366
18	3	6	0.399	−0.331	−0.154
19	4	1	0.175	0.090	−0.414
20	4	2	0.394	0.137	−0.036
21	4	3	0.035	0.562	0.140
22	4	4	0.619	0.113	0.213
23	4	5	0.363	−0.048	0.170
24	4	6	0.272	0.043	−0.448
25	5	1	0.588	−0.124	−0.144
26	5	2	0.338	−0.122	−0.243
27	5	3	0.151	0.648	−0.070
28	5	4	0.296	0.064	−0.463
29	5	5	0.289	−0.438	0.336
30	5	6	0.683	−0.235	−0.153
31	6	1	0.355	0.038	−0.354
32	6	2	0.528	0.018	0.430
33	6	3	0.428	0.345	0.463
34	6	4	0.424	−0.028	0.026
35	6	5	0.230	−0.182	0.342
36	6	6	0.431	0.109	−0.106
Overall% of explained variance			**17.1%**	**8.7%**	**6.9%**

Appendix 1b. Exploratory factor analysis on the data from the visual SL task in Experiment 3, condition 2. Included here are loadings of all trials on the main three extracted factors. Numbers under 'target' and 'foil' correspond to the presented stimuli in Table 2 above.

Serial trial no.	Target word	Foil	Factor 1 loading	Factor 2 loading	Factor 3 loading
1	1	1	0.001	0.396	0.069
2	1	2	−0.287	0.538	0.212
3	1	3	−0.262	0.213	0.385
4	1	4	0.043	0.151	0.313
5	1	5	0.238	0.282	0.109
6	1	6	0.463	−0.482	0.053
7	2	1	0.365	0.414	−0.037
8	2	2	0.220	0.046	0.517
9	2	3	0.341	0.143	0.266
10	2	4	0.450	−0.038	0.362
11	2	5	0.587	−0.248	0.096
12	2	6	0.557	−0.068	0.228
13	3	1	0.486	0.106	0.025
14	3	2	0.411	−0.009	−0.084
15	3	3	0.356	−0.259	0.136
16	3	4	0.315	−0.417	0.439
17	3	5	0.241	0.010	−0.028
18	3	6	0.447	−0.342	−0.221
19	4	1	0.057	0.365	−0.112
20	4	2	0.108	0.322	−0.415
21	4	3	0.234	−0.273	−0.359
22	4	4	0.248	.261	0.281
23	4	5	0.168	0.095	−0.160
24	4	6	0.420	−0.346	−0.214
25	5	1	0.138	0.284	−0.532
26	5	2	0.191	0.515	−0.192
27	5	3	0.355	0.124	−0.216
28	5	4	0.400	0.444	−0.077
29	5	5	0.550	0.167	−0.333
30	5	6	0.335	−0.002	−0.469
31	6	1	0.413	0.281	0.007
32	6	2	0.396	0.411	0.297
33	6	3	0.428	0.112	0.402
34	6	4	0.624	−0.152	0.045
35	6	5	0.541	0.284	0.011
36	6	6	0.704	−0.123	−0.120
Overall% of explained variance			**14.6%**	**8.1%**	**7.1%**

Appendix 2a. Exploratory factor analysis on the data from the auditory SL task in Experiment 3, condition 1. Included here are loadings of all trials on the main three extracted factors.

Serial trial no.	Target word	Foil	Factor 1 loading	Factor 2 loading	Factor 3 loading
1	lenamo	lerifa	0.245	0.205	0.227
2	lenamo	minade	0.442	0.199	0.295
3	lenamo	nukaro	0.001	−0.262	0.047
4	lenamo	pabose	0.568	−0.096	0.123
5	lenamo	savogi	0.282	−0.314	0.404
6	lenamo	tilumo	0.113	0.221	0.104
7	mivofa	lerifa	0.376	0.377	0.063
8	mivofa	minade	0.549	−0.358	0.123
9	mivofa	nukaro	0.248	−0.043	0.427
10	mivofa	pabose	0.225	0.197	0.095
11	mivofa	savogi	0.177	0.080	0.277
12	mivofa	tilumo	0.360	0.259	0.157

13	nubogi	lerifa	−0.187	0.695	0.212
14	nubogi	minade	−0.106	0.400	0.230
15	nubogi	nukaro	−0.326	0.011	0.529
16	nubogi	pabose	−0.060	0.383	0.277
17	nubogi	savogi	−0.019	0.071	0.583
18	nubogi	tilumo	−0.270	0.376	0.028
19	paluro	lerifa	0.222	0.537	0.335
20	paluro	minade	0.466	−0.245	−0.015
21	paluro	nukaro	0.116	0.073	0.025
22	paluro	pabose	0.475	−0.094	0.071
23	paluro	savogi	0.222	−0.309	0.371
24	paluro	tilumo	0.352	0.347	0.258
25	saride	lerifa	0.478	−0.023	−0.250
26	saride	minade	0.001	−0.321	−0.172
27	saride	nukaro	−0.481	0.106	−0.174
28	saride	pabose	0.016	−0.371	0.452
29	saride	savogi	0.228	−0.277	0.134
30	saride	tilumo	−0.269	0.153	−0.169
31	tikase	lerifa	−0.482	−0.015	0.288
32	tikase	minade	−0.285	−0.434	0.297
33	tikase	nukaro	−0.450	0.204	0.302
34	tikase	pabose	−0.252	−0.265	0.494
35	tikase	savogi	−0.479	−0.254	0.387
36	tikase	tilumo	−0.341	−0.345	0.256
Overall% of explained variance			**10.6%**	**8.4%**	**7.9**

Appendix 2b. Exploratory factor analysis on the data from the auditory SL task in Experiment 3, condition 2. Included here are loadings of all trials on the main three extracted factors.

Serial trial no.	Target word	Foil	Factor 1 loading	Factor 2 loading	Factor 3 loading
1	dukeva	dulize	0.404	−0.048	0.038
2	dukeva	kugabe	.281	−0.082	0.057
3	dukeva	nitomu	0.505	−0.259	0.427
4	dukeva	nodiva	0.163	0.037	0.526
5	dukeva	sokeba	−0.014	0.375	−0.068
6	dukeva	vugota	0.320	−0.185	0.046
7	kutoze	dulize	0.023	0.398	0.311
8	kutoze	kugabe	0.317	0.052	0.268
9	kutoze	nitomu	0.074	−0.274	0.467
10	kutoze	nodiva	0.005	0.057	0.401
11	kutoze	sokeba	−0.171	0.117	0.140
12	kutoze	vugota	0.235	0.047	−0.312
13	nigobe	dulize	0.472	0.153	0.206
14	nigobe	kugabe	0.473	0.059	−0.143
15	nigobe	nitomu	0.050	−0.174	0.438
16	nigobe	nodiva	−0.319	0.312	0.560
17	nigobe	sokeba	0.238	0.598	−0.053
18	nigobe	vugota	0.022	−0.034	−0.288
19	nolita	dulize	−0.033	0.672	−0.006
20	nolita	kugabe	0.144	0.614	−0.045
21	nolita	nitomu	0.501	0.045	−0.316
22	nolita	nodiva	0.531	0.358	0.012
23	nolita	sokeba	−0.044	0.532	0.014
24	nolita	vugota	0.514	0.006	−0.394
25	sogamu	dulize	−0.263	0.363	−0.174
26	sogamu	kugabe	−0.224	0.450	0.081
27	sogamu	nitomu	0.007	−0.054	0.403
28	sogamu	nodiva	−0.276	0.100	0.220
29	sogamu	sokeba	−0.599	0.197	0.048
30	sogamu	vugota	0.064	0.077	−0.113
31	vudiba	dulize	0.327	0.310	0.227

32	vudiba	kugabe	0.575	0.043	−0.229
33	vudiba	nitomu	0.436	0.010	0.111
34	vudiba	nodiva	0.239	0.149	0.485
35	vudiba	sokeba	−0.172	0.293	−0.315
36	vudiba	vugota	0.322	0.166	0.110
Overall% of explained variance			**10.1%**	**8%**	**7.7%**

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language, 62*(3), 311–331. http://dx.doi.org/10.1016/j.jml.2009.11.007.

Arciuli, J. (2017). The multi-component nature of statistical learning. Philosophical *Transactions of the Royal Society of London. Series B, Biological Sciences*, 372. http://dx.doi.org/10.1098/rstb.2016.0058.

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science, 14*, 464–473. http://dx.doi.org/10.1111/j.1467-7687.2009.00937.x.

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*, 286–304. http://dx.doi.org/10.1111/j.1551-6709.2011.01200.x.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–347. http://dx.doi.org/10.1097/AUD.0b013e31821473f7.

Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic Bulletin & Review, 23*(4), 1250–1256.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes. *Psychological Science, 19*(7), 678–685. http://dx.doi.org/10.1111/j.1467-9280.2008.02142.x.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition, 121*, 127–132. http://dx.doi.org/10.1016/j.cognition.2011.06.010.

Campbell, K. L., Zimerman, S., Healey, M. K., Lee, M. M. S., & Hasher, L. (2012). Age differences in visual statistical learning. *Psychology and Aging, 27*(3), 50–56. http://dx.doi.org/10.1037/a0026780.

Christiansen, M. H., & Curtin, S. (1999). Transfer of learning: Rule acquisition or statistical learning? *Trends in Cognitive Sciences, 3*(8), 289–290.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2000). A connectionist single-mechanism account of rule-like behavior in infancy. In *Proceedings of the 22nd meeting of the cognitive science society* (pp. 83–86). Cognitive Science Society.

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*(1), 24–39. http://dx.doi.org/10.1037/0278-7393.31.1.24.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. http://dx.doi.org/10.1007/BF02310555.

Cunillera, T., Laine, M., Camara, E., & Rodriguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language, 63*, 295–305. http://dx.doi.org/10.1016/j.jml.2010.05.003.

Diedenhofen, B., & Musch, J. (2016). Cocron: A web interface and R package for the statistical comparison of cronbach's alpha coefficients. *International Journal of Internet Science, 11*(1), 51–60.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology, 92*, 37–64. http://dx.doi.org/10.1016/j.cogpsych.2016.11.004.

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. http://dx.doi.org/10.1016/j.jml.2008.10.003.

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. S. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra, 2*(1), 14. http://dx.doi.org/10.1525/collabra.41.

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition, 108*(2), 477–499. http://dx.doi.org/10.1016/j.cognition.2008.04.002.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*, 499–504. http://dx.doi.org/10.1111/1467-9280.00392.

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*(3), 117–125. http://dx.doi.org/10.1016/j.tics.2014.12.010.

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science, 24*(7), 1243–1252. http://dx.doi.org/10.1177/0956797612472207.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science, 33*(6), 1087–1116. http://dx.doi.org/10.1111/j.1551-6709.2009.01041.x.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review, 16*, 486–490. http://dx.doi.org/10.3758/PBR.16.3.486.

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin & Review, 20*, 1161–1169. http://dx.doi.org/10.3758/s13423-013-0458-4.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*(1), 21–54. http://dx.doi.org/10.1016/j.cognition.2009.03.008.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431–436. http://dx.doi.org/10.1111/1467-9280.00476.

Karuza, E. A., Farmer, T. A., Fine, A. B., Smith, F. X., & Jaeger, T. F. (2014). On-line Measures of Prediction in a Self-Paced Statistical Learning Task. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 725–730).

Karuza, E. A., Li, P., Weiss, D. J., Bulgarelli, F., Zinszer, B. D., & Aslin, R. N. (2016). Sampling over nonuniform distributions: A neural efficiency account of the primacy effect in statistical learning. *Journal of Cognitive Neuroscience, 28*(10), 1484–1500. http://dx.doi.org/10.1162/jocn_a_00990.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development, 87*(1), 184–193. http://dx.doi.org/10.1111/cdev.12461.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), B35–B42. http://dx.doi.org/10.1016/S0010-0277(02)00004-5.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science, 14*(6), 1323–1329. http://dx.doi.org/10.1111/j.1467-7687.2011.01079.x.

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition, 122*(2), 241–246. http://dx.doi.org/10.1016/j.cognition.2011.10.007.

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition, 1–25*. http://dx.doi.org/10.3758/s13421-011-0074-3.

Milne, A. E., Petkov, C. I., & Wilson, B. (2017). Auditory and visual sequence learning in humans and monkeys using an artificial grammar learning paradigm. *Neuroscience*. http://dx.doi.org/10.1016/j.neuroscience.2017.06.059.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162. http://dx.doi.org/10.1016/S0010-0285(03)00128-2.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language, 53*(2), 225–237. http://dx.doi.org/10.1016/j.jml.2005.02.011.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36*(7), 1299–1305. http://dx.doi.org/10.3758/MC.36.7.1299.

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*(4), 807–818. http://dx.doi.org/10.1016/j.jml.2012.02.010.

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta Psychologica, 149*, 1–8. http://dx.doi.org/10.1016/j.actpsy.2014.01.015.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246–263. http://dx.doi.org/10.1006/jmla.1998.2576.

Poulin-Charronnat, B., Perruchet, P., Tillmann, B., & Peereman, R. (2016). Familiar units prevail over statistical cues in word segmentation. *Psychological Research Psychologische Forschung, 1–14*. http://dx.doi.org/10.1007/s00426-016-0793-y.

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*. http://dx.doi.org/10.1111/desc.12593.

Revelle, W. (2016). Psych: Procedures for personality and psychological research. *R Package,* 1–358. Retrieved from <http://personality-project.org/r/psych-manual.pdf>.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. http://dx.doi.org/10.1126/science.274.5294.1926.

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. http://dx.doi.org/10.1006/jmla.1996.0032.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*(2), 101–105. http://dx.doi.org/10.1111/j.1467-9280.1997.

tb00690.x.
Schapiro, A. C., Gregory, E., & Landau, B. (2014). The necessity of the medial-temporal lobe for statistical learning. *Journal of Cognitive Neuroscience, 26*, 1736–1747.
Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience, 26*(8), 1736–1747. http://dx.doi.org/10.1162/jocn_a_00578.
Siegelman, N., Bogaerts, L., & Frost, R. (2016). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 1–15*. http://dx.doi.org/10.3758/s13428-016-0719-z.
Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120. http://dx.doi.org/10.1016/j.jml.2015.02.001.
Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103. http://dx.doi.org/10.1207/S15327752JPA8001_18.
Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin, 139*, 792–814. http://dx.doi.org/10.1037/a0030801.
Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*(4), 706–716. http://dx.doi.org/10.1037/0012-1649.39.4.706.
Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3*(1), 1–42. http://dx.doi.org/10.1207/s15473341lld0301_1.
Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology-General, 134*(4), 552–564. http://dx.doi.org/10.1037/0096-3445.134.4.552.