LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Close Encounters of the Word Kind: Attested Distributional Information Boosts Statistical Learning

Katja Stärk [a], Evan Kidd [a,b,c] and Rebecca L. A. Frost [a,d]

[a]Max Planck Institute for Psycholinguistics [b]The Australian National University [c]ARC Centre of Excellence for the Dynamics of Language [d]Edge Hill University

CRediT author statement – **Katja Stärk**: conceptualization (equal); data curation (lead); formal analysis (lead); funding acquisition (equal); investigation (lead); methodology (equal); project administration (equal); validation (lead); visualization (lead); writing – original draft preparation (lead); writing – review & editing (lead). **Evan Kidd**: conceptualization (equal); funding acquisition (equal); methodology (equal); administration (equal); supervision (equal); writing – original draft preparation (supporting); writing – review & editing (supporting). **Rebecca L. A. Frost**: conceptualization (equal); funding acquisition (equal); methodology (equal); administration (equal); supervision (equal); writing – original draft preparation (supporting); writing – review & editing (supporting).

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

Katja Stärk is supported by a PhD studentship from the Language Development Department, Max Planck Institute for Psycholinguistics, and Evan Kidd is supported by the Australian Research Council (CE140100041). We have no known conflict of interest to disclose. We thank Emma Marsden, Theres Grüter, and four anonymous reviewers, as well as the members of the Language Development Department at the Max Planck Institute for Psycholinguistics, for their insightful comments on this work. Special thanks go to Andrew Jessop for his helpful guidance in R programming and statistics. Thanks also to Julia Egger for providing the Python script to combine text files, to Greta Kaufeld for recording the stimuli, and to Nico Pani and Caroline De Becker for piloting the study and transcribing parts of the data for the reliability analysis.

Correspondence concerning this article should be addressed to Katja Stärk, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. E-mail: katja.staerk@mpi.nl

The handling editor for this article was Theres Grüter.

DOI: 10.1111/lang.12523

Abstract: Statistical learning, the ability to extract regularities from input (e.g., in language), is likely supported by learners' prior expectations about how component units co-occur. In this study, we investigated how adults' prior experience with sublexical regularities in their native language influences performance on an empirical language learning task. Forty German-speaking adults completed a speech repetition task in which they repeated eight-syllable sequences from two experimental languages: one containing disyllabic words comprised of frequently occurring German syllable transitions (naturalistic words) and the other containing words made from unattested syllable transitions (non-naturalistic words). The participants demonstrated learning from both naturalistic and non-naturalistic stimuli. However, learning was superior for the naturalistic sequences, indicating that the participants had used their existing distributional knowledge of German to extract the naturalistic words faster and more accurately than the non-naturalistic words. This finding supports theories of statistical learning as a form of chunking, whereby frequently co-occurring units become entrenched in long-term memory.

Keywords statistical learning; serial recall; incremental learning; long-term memory; entrenchment

Introduction

Humans are exquisitely sensitive to the regularities in their environment. Statistical learning (SL), the ability to draw on these regularities, is hypothesized to underlie learning across all sensory domains. Although it is indisputable that humans are capable of SL (which might rely upon multiple interacting mechanisms, see Frost et al., 2015), the totality of the parameters influencing SL are still yet to be mapped out. In our study, we examined the degree to which SL of linguistic stimuli is influenced by prior knowledge of attested syllable transitions present in natural language. That is, we asked whether and how adults' prior experience with the sublexical regularities in their native language in the form of syllable bigrams would permeate into the laboratory, such that it would enhance the adults' performance on an empirical language learning task when the distributional properties of the to-be-learned material aligned with those of the natural language.

In a canonical auditory SL task using linguistic stimuli, participants listen to a stream of speech that contains to-be-discovered words that are defined by statistical regularities (e.g., Saffran et al., 1996). The discovery of the statistical segmentation effect heralded great promise for non-nativist approaches to language acquisition because it suggested the existence of a powerful learning mechanism (or mechanisms) that can induce structure from the input and thus questioned the need to postulate innately specified language-specific

knowledge. That even very young infants are capable of SL is not controversial; however, the parameters that influence the process are still not well understood. This is partly due to the fact that much of the research on the topic has been conducted independently from other fields in cognitive psychology (Frost et al., 2019), such that connections to older disciplines concerned with learning and memory have not always been made. Yet, any task concerning learning of linguistic stimuli should be expected to conform to well-known properties of verbal memory, with SL being no exception (Isbilen et al., 2020; Vlach & DeBrock, 2017; Vlach & Sandhofer, 2014).

Since as far back as Ebbinghaus (1885, 1913), researchers have known that verbal learning is most effective when learners build upon prior experience. Accordingly, if researchers are to take the results of SL research to be ecologically valid, they should not expect participants to come into the laboratory without prior implicit assumptions about how linguistic stimuli like phonemes and syllables are ordered (Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Siegelman et al., 2018) and should instead expect participants to learn best when the target language is consistent with those assumptions. Such historically-contingent and in many instances top-down influences on performance suggest that the output of SL shapes future learning.

Background Literature

A growing number of studies have shown that prior knowledge and expectations derived from a speaker's native language shape subsequent SL in a number of ways. This process begins very early. For example, Lew-Williams and Saffran (2012) found that infants' statistical segmentation of novel words from continuous speech was guided by their experience with words of the same versus a different length in a pretraining phase such that segmentation was only possible when words were the same length in both exposure phases. Similarly, research has revealed a significant benefit of starting small during incremental SL, with learners bootstrapping upon initial experience with simpler structures. Zettersten et al. (2020) demonstrated that adults' prior experience with a simplified nonadjacent dependency-learning task boosted later learning of a more complex instantiation of the same structure (see also Lany & Gómez, 2008, for similar findings with infants). In related work, Lai and Poletiek (2011) found that exposure to simple AB dependencies helped subsequent learning of longer, more complex strings containing center-embeddings. Together, these findings provide converging evidence that prior experience scaffolds for future learning of related material.

Importantly, similar transfer effects have been found to emerge through experience with natural as well as artificial languages and across different learning contexts. For instance, Potter et al. (2017) documented a language-experience effect in novice learners of Mandarin after just two semesters of study. In Potter et al.'s study, participants completed a SL task in which the artificial language overlapped with Mandarin insofar as it was tonal in nature (see also Wang & Saffran, 2014) to see whether participants' experience with related material would impact learning. Participants completed the task at two time points separated by an interim learning period of six months. Although participants' performance was initially at chance, there were significant improvements at Time 2, with participants achieving accuracy scores of 66% on a two-alternative forced-choice (2AFC) segmentation test, indicating that participants' SL performance had been critically shaped by their experience with relevant linguistic input. Non-Mandarin-learning controls exhibited no such improvements, performing at chance on both occasions.

Other studies have investigated how statistical distributions in naturalistic language constrain SL in laboratory settings, with a large focus on phonotactic probabilities (for reviews of how phonotactics impact on early acquisition, see Johnson, 2016; Jusczyk, 2002). For instance, Finn and Hudson Kam (2008) showed that participants could only successfully segment statistically defined novel words from continuous speech when the words contained syllables that followed phonotactic constraints of English (see also Toro et al., 2011). Mersad and Nazzi (2011) showed that the presence of words containing high phonotactic probability served as anchors that successfully aided segmentation compared to a condition in which all words had a uniform but lower phonotactic probability. Dal Ben et al. (2021) replicated this latter effect using a more narrowly defined difference in phonotactic probability across experimental conditions. Overall, these studies provide strong evidence for the suggestion that fine-grained features of natural language, in this case phonotactic probability, shape participants' subsequent expectations about how their input is shaped. This is consistent with results reported by Siegelman et al. (2018), who found that performance on an auditory SL segmentation task was predicted by post hoc ratings of how word-like test items and foils were.

These findings provide converging support for the notion that prior experience can shape future learning at multiple levels of description, boosting performance when the properties of the input align. Building on this, Elazar et al. (2022)[1] investigated the specific hypothesis that entrenched memory traces for syllable co-occurrences in natural language boost SL. They tested two groups of Spanish-speaking participants on an auditory SL task. One group listened

to a Spanish-like speech stream in which transitional probabilities (TPs) of the experimental words were highly attested in Spanish while the other group listened to a Spanish-unlike speech stream in which TPs of the experimental words were rarely attested in Spanish. Participants were tested on a lexical decision task for experimental words and respective Spanish-like or Spanish-unlike foils. Elazar et al. found that participants in the Spanish-like condition were better at accepting words than participants in the Spanish-unlike condition, indicating that participants' prior knowledge of Spanish syllable trigrams facilitated their SL. Furthermore, participants in the Spanish-like condition were worse at rejecting (Spanish-like) foils than participants in the Spanish-unlike condition were at rejecting (Spanish-unlike) foils, suggesting that participants' knowledge of Spanish also (mis)led them to accept familiar foils. Overall, the results suggest that participants indeed entered the experiment with entrenched memory traces for syllable co-occurrences on which they drew to process and learn new input.

The Present Study

In our study, on which we worked in parallel to Elazar et al.'s (2022) study, we tested the almost identical hypothesis that entrenched memory traces for syllable bigrams in natural language boost SL. However, Elazar et al. (2022) used a between-participants design following the typical exposure-phase–test-phase structure, whereas we used a within-participants design using verbal repetition. This within-participant design allowed for a more stringent test of the entrenchment hypothesis because differences between conditions could not be attributed to differences between participants, in addition to allowing us to track the emergence of learning across the course of the experiment. In using verbal repetition, we built upon recent developments in the measurement of SL that have been inspired by the verbal learning literature. Participants' recall on verbal tests of short-term memory has been shown to be both sensitive to newly learned material (e.g., Majerus et al., 2004) and mediated by their long-term lexical knowledge (e.g., Kowialiewski & Majerus, 2018; Majerus et al., 2012; Majerus et al., 2004). In recent work building upon Majerus et al. (2004), Isbilen et al. (2020) investigated the utility of verbal recall as a measure of auditory SL in a triplet segmentation task. Isbilen et al. showed that, after familiarization with continuous speech, adult participants were better able to repeat syllable sequences that followed the statistical distribution of the familiarization stream than they were to repeat random and unattested syllable sequences (for similar results from children, see Kidd et al., 2020). In some cases, performance was predicted by distributional statistics derived from

spoken and written corpora. Isbilen et al. suggested that the results were consistent with chunking models of SL (e.g., Christiansen & Chater, 2016; Jones, 2012; Perruchet & Vinter, 1998; Robinet et al., 2011) in which the repetition of syllable sequences creates word-like phonological units via their association strength; specifically, their high TPs.

These processes and their explanation seem functionally equivalent to another effect in the literature—the Hebb repetition effect (Hebb, 1961; see also Page & Norris, 2009; Smalle et al., 2016; Szmalec et al., 2012). However, the one difference between auditory SL tasks and Hebbian learning tasks is that, although SL tasks typically measure the outcome of learning following familiarization, Hebbian learning tasks track learning of sequence regularities across time. This is an important gap in SL research, with researchers not yet knowing how learning proceeds during familiarization. The evidence that exists has suggested that learners gradually come to recognize structured sequences as containing higher level chunks over the course of exposure, suggesting that learners engage in the dual processes of (a) binding/chunking adjacent syllables together and (b) storing them in long-term memory (Batterink & Paller, 2017).

In our study, we used a sequence-repetition method common in Hebbian learning studies to also investigate how existing knowledge of sublexical regularities influences the trajectory of SL over time. Our article makes two contributions to the literature: (a) we report detailed corpus data on syllable transitions in German, and (b) we determine how these attested transitions contribute to SL across the course of learning. Thus, building on previous investigations of the effect of prior knowledge on SL (e.g., Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Siegelman et al., 2018; Toro et al., 2011), we examined how knowledge of the statistical properties of participants' native language—focusing on syllable bigrams—influences subsequent processing and learning of an artificial language that is built with those properties in mind. We extracted the TPs between syllable pairs in natural German and used this information to create artificial language sequences containing words that were either based on the natural German TPs (i.e., naturalistic sequences) or not (i.e., non-naturalistic sequences), examining learning of these sequences relative to scrambled foils.

Under the assumption that SL for language involves the tracking and subsequent long-term registration of distributional information, we hypothesized that learners would use their existing distributional knowledge of German to shape their processing of new input. To test this hypothesis, we measured learning using a speech production task in which the participants repeated either

unstructured sequences of random syllable combinations (foils) or structured sequences containing novel words—with these words either adhering to German syllable distribution (i.e., naturalistic sequences), or not adhering to German syllable distribution (i.e., non-naturalistic sequences). We predicted that, overall, participants' repetition (and therefore learning) of the structured sequences would be better than their repetition of the foils, but that participants' repetition of the naturalistic sequences would be better than their repetition of the non-naturalistic sequences. An advantage of our method was that, in contrast to past research measuring learning via 2AFC and repetition after familiarization, it enabled us to track learning across the three conditions across the course of the experiment. We also predicted that, over time, participants' repetitions would improve for both types of structured sequences. Importantly, we expected to see the strongest improvements for naturalistic sequences and predicted that performance would improve more rapidly for naturalistic than for non-naturalistic sequences because naturalistic sequences better aligned with German syllable distributions.

Method

All materials, data, analyses, and results (Stärk et al., 2022b) for this article are openly available via OSF (https://osf.io/4dsmy); the results of the experiment testing the validity of the stimuli (Stärk et al., 2022c) can also be accessed via OSF (https://osf.io/p9fcm).

Participants

Forty native German-speaking adults (28 self-identified female, 12 self-identified male; $M_{age} = 23.9$ years, $SD = 5.58$) without any known hearing, speech, or language disorders participated in the experiment. The participants registered via the Max Planck Institute's internal database; we made additional announcements at Radboud University and on social media, which also allowed participants to register via email. The sample size of 40 participants was informed by a power analysis conducted in the software R (R Core Team, 2020) using the package simr 1.0.5 (Green & MacLeod, 2016). We based the simulations on data collected by Isbilen et al. (2017), who had compared two conditions similar to our non-naturalistic and unstructured foil sequences in a serial recall task following an exposure phase. Our simulations indicated that a sample of 16 participants would be sufficient to detect an effect size of a –0.1 syllable recall difference between naturalistic and non-naturalistic sequences as well as between non-naturalistic and foil sequences during the later stages of our experiment, which is comparable to the test phase in Isbilen

et al.'s study (for more details, see the analysis folder of the project's OSF page at https://osf.io/4dsmy). We increased the sample size to 40 because the participants in our experiment were exposed to multiple experimental languages while performing the serial recall task (i.e., without prior exposure phase), which would decrease the effect and also make the model more complex (because we included the additional variable block, which was not present in Isbilen et al.'s study). We decided to not perform an analysis modeling our entire experiment (including block) because this would have entailed a considerably more complex simulation that would have been based purely on our own intuitions rather than on previous data.

The study was approved by the Ethical Committee of the Faculty of Social Sciences, Radboud University, and was carried out in accordance with the World Medical Association Declaration of Helsinki. All participants gave written informed consent prior to their participation in the study. They were free to withdraw at any time and were compensated (€8) upon completing the 45-minute session.

Design

We employed a serial repetition task based on studies of the Hebb repetition effect (Hebb, 1961; Page & Norris, 2009), which required the participants to repeat sequences of syllables aloud, with these repetitions then being scored for accuracy. The study had a within-participants design, with all the participants receiving three different types of sequences: (a) naturalistic sequences, (b) non-naturalistic sequences, and (c) unstructured foils. The naturalistic and non-naturalistic sequences were structured, with each containing four disyllabic experimental words, whereas foils were unstructured, containing the same syllables as the structured sequences but in a scrambled order.

Materials

Corpus Analysis

We created the speech stimuli from a pool of 12 German syllables (*fa*, *ge*, *gei*, *mi*, *mo*, *nu*, *pa*, *sa*, *su*, *ti*, *ver*, *zu*) obtained from a corpus analysis of the 1,000 most frequent German words in the CHILDES database (MacWhinney, 2000),[2] which corresponded to over three million word tokens. We chose to draw our materials from child-directed language for two reasons. First, because words that are highly frequent in child-directed language will also have an early age-of-acquisition, we logically deduced that these words would have sublexical transitions (i.e., bigrams) that would have the greatest likelihood of being entrenched. Second, this study was part of a larger project that tested the

Table 1 Syllable frequencies, pair frequencies, and forward and backward transitional probabilities of the stimuli derived from the corpus analysis

	Frequency			Transitional probability	
Pair	Syllable 1	Syllable 2	Pair	Forward	Backward
mi nu	6,472	454	454	0.070	1.000
pa gei	46,359	368	368	0.008	1.000
ver su	14,010	344	344	0.025	1.000
ge fa	1,839,597	3,133	1,586	0.001	0.506
zu sa	14,460	4,670	1,994	0.138	0.427
mo ti	1,748	2,467	525	0.300	0.213

effects under investigation in developmental populations. We chose the syllables from syllable pairs (i.e., bigrams) occurring with high within-word backward TPs, relying on backward TPs because a corpus analysis of child-directed speech by Stärk et al. (2022a) showed that backward TPs are significantly more reliable cues to wordhood than forward TPs in German speech (for a similar cross-linguistic analysis, see Saksida et al., 2017).

We then used the syllables to form 12 disyllabic "words": six words for each of the two structured sequence types (naturalistic: *gefa*, *minu*, *moti*, *pagei*, *versu*, *zusa*; non-naturalistic: *fazu*, *geimi*, *nuver*, *samo*, *suge*, *tipa*). As summarized in Table 1, the extracted bigrams yielded the six naturalistic words in which the two syllables co-occurred with relatively high backward TPs in natural German speech but importantly were not recognizable alone as words (TP $> .20$, $M_{TP} = .69$, *range* $= .21$–1.00).[3] To create the non-naturalistic words, we concatenated the same 12 syllables in a different order, such that their syllable pairs did not co-occur in natural German (TP $= .00$). Each syllable occurred once in each set of words, and we counterbalanced the position of syllables within words such that, if a syllable appeared word-initially in the naturalistic set of words, it was word-final in the non-naturalistic set, and vice versa. For the unstructured foil sequences, we scrambled the syllables, such that these sequences contained no learnable regularities. We carefully constructed the foils to avoid inadvertently creating words from both German and the experimental languages. Because all three sequence types comprised the same 12 syllables, the frequencies of the syllables in natural German presented in Table 1 applied to all conditions. However, the non-naturalistic words comprised syllable pairs which did not occur in our corpus sample of natural German (i.e., their pair frequencies as well as their forward and backward TPs were

0). Likewise, the unstructured foils did not comprise any patterns found in the corpus.

Our design involved the explicit assumption that high TPs are more word-like and, thus, that the participants would require less exposure to chunk adjacent syllables into words. The implicit assumption of our sequence repetition method was that these transitions would thus be easier to repeat. In order to collect independent evidence in support of the explicit assumption that the naturalistic words would be more word-like, we conducted a separate experiment in which we asked German-speaking participants to select our naturalistic or non-naturalistic words for wordiness in comparison to foils in a 2AFC task without familiarization (i.e., the participants had no prior training on the words). The participants successfully identified the naturalistic words at above chance levels in comparison to foils but did not do so for the non-naturalistic words. These results were consistent with the argument that our naturalistic words, when presented in isolation, were more identifiable as German-like than our non-naturalistic words (for full details, see Appendix S1 in the Supporting Information online).

Stimuli Characteristics

The stimuli were recorded by a female native speaker of German, who recorded individual unstressed syllables in isolation. We adjusted the syllable recordings using the sound editing program Audacity (Audacity Team, 2018) to ensure uniformity in length, resulting in an average syllable duration of 377 ms (*range* = 352–416).

Within the context of the experiment, each structured sequence type contained perfect within-word TPs (structured sequences: within-word TPs = 1.00, between-word TPs $\leq$ .25; compared to unstructured sequences where TPs between all syllables were generally low, with TPs $\leq$ .125). Note, however, that participants were tested on all three sequence types. Thus, across the whole experiment, accounting for the repeated use of syllables across each type of sequence, within-word TPs for both structured sequences were .33, and TPs for all other syllable pairs were less than or equal to .125.

Syllables were combined into 72 sequences, 24 of each sequence type. Each sequence was eight syllables long, which equated to four experimental words (i.e., bigrams). Within sequences, each syllable was followed by 500 ms of silence, and the final syllable of a sequence was followed by a beep to indicate the beginning of the repetition portion of the trial. Because the syllables had an inter-stimulus interval of 500 ms, we emphasize that our study was not a segmentation task in the classical sense. Rather, our choice of method

allowed us to determine (a) whether attested syllable bigrams are more naturally grouped during recall and (b) how this attested knowledge influences learning incrementally across time. In order to track the participants' incremental learning, we divided the experiment into 12 blocks of six sequences, with each block containing two sequences of each type. Within each block, sequences were presented pseudo-randomly, with no direct repetition of a particular sequence type. Across the whole experiment, each word occurred 16 times in total, with words appearing equally often in each position within a sequence (for more information on the stimuli and their creation, see the materials folder on the project's OSF page at https://osf.io/4dsmy).

Procedure

We sent the participants an informed consent document one day prior to the day that the study took place. Upon arrival in the laboratory, they were reminded of the task instructions and were told that the study was to investigate adults' repetition of language, but no mention was made of the learnable patterns contained within the input. The participants completed the study in isolation in a sound-attenuated booth, with sequences being played over closed-cup headphones using the software Presentation (Neurobehavioral Systems, 2014). The participants repeated the sequences into a microphone, and these were recorded by the computer for offline coding.

Before the experiment began, the participants first received three (unstructured) practice sequences that were six syllables long, comprising a different scrambled set of syllables (*ba*, *fun*, *gi*, *re*, *se*, *to*). After completing the practice sequences, the participants proceeded to the main experiment. In each trial, the participants heard a sequence of eight syllables followed by a beep (see Figure 1). Upon hearing the beep, they were required to repeat the sequences as best they could. At the halfway point, the participants were given the opportunity to take an optional break. At the end of the session, they were debriefed and paid for their time.

Data Preparation

To prepare the data for our analyses, we first transcribed the recordings of the participants' verbal responses. All responses were transcribed by the experimenter, and two naïve coders each transcribed 10% of the recordings (i.e., data for four participants) for the purpose of performing reliability checks. Inter-transcriber reliability analyses revealed strong reliability between transcribers using the more conservative interpretation of the kappa statistic suggested by McHugh (2012): syllable-level observed agreement = 83.0%, $\kappa = .87$, 95%

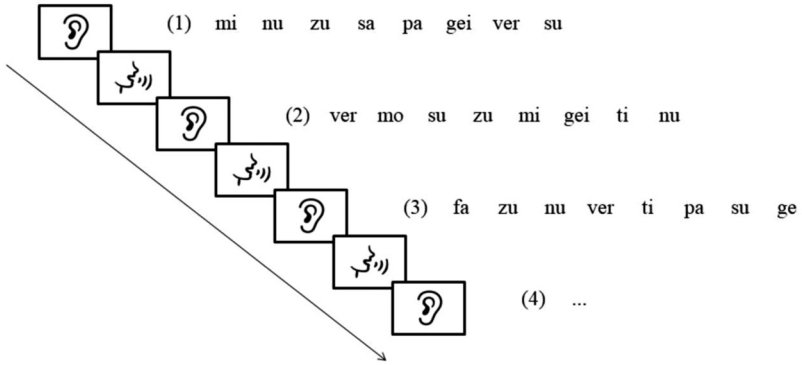

Figure 1 Three example experimental sequences. On each trial, participants listened to a sequence and then repeated it. (1) = one naturalistic sequence; (2) = one unstructured foil sequence; (3) = one non-naturalistic sequence.

CIs [.84, .89]; bigram-level observed agreement = 87.2%, κ = .88, 95% CI [.84, .92].

We coded the accuracy of participants' responses sequence-by-sequence, comparing the verbal response against the sequence that the participants had heard. We computed scores at the syllable level and at the bigram level. At the syllable level, the participants received 1 point for each syllable repeated correctly in the correct position (for a maximum of 8 points per sequence). At the bigram level, the participants received 1 point for each bigram (i.e., syllable pair) repeated correctly in the correct position (for a maximum of 4 points per sequence). A bigram denoted an experimental word in the structured sequences. The participants' performance at this level, therefore, provided crucial information about whether they had recalled sequences better because of learning the experimental words, rather than indirectly assessing the learning solely at the syllable level. In the unstructured sequences, the four bigrams per sequence were the random syllable pairs in Positions 1 and 2, 3 and 4, 5 and 6, and 7 and 8 (with different syllable combinations in each position across each sequence). Table 2 illustrates how we applied the coding scheme to potential repetitions by the participants.

This strict coding scheme was conservative in the sense that it required the participants to recall elements in the correct position and thus did not give any credit for syllables recalled in the correct order after only one syllable was missed or added in the repetition (e.g., as in the "B C D E F G H A" and "A X B C D" cases in Table 2). We also used two further coding schemes, including a

Table 2 Example scorings of participants' repetitions of the sequence "A B C D E F G H" where each letter represents one syllable

Repetition	Syllable score	Bigram score
A B C D F H –	4 (A B C D)	2 (AB CD)
A X B C D –	1 (A)	0
A B –	2 (A B)	1 (AB)
X Y C D X Z G H	4 (C D G H)	2 (CD GH)
B C D E F G H A	0	0

serial order coding scheme based on Isbilen et al.'s (2017) study, which relaxed the strict positional requirement and which were thus more lenient and gave more credit to the participants. However, because all three analyses converged in the same direction, we have chosen to report only the most conservative scheme here. The analyses for all three coding schemes can be found in the analysis folder on the project's OSF page at https://osf.io/4dsmy.

Results

The aims of our analyses were twofold: (a) to examine performance on each type of sequence and (b) to examine the time course of learning. We predicted that the participants would recall naturalistic sequences better than non-naturalistic sequences and non-naturalistic sequences better than unstructured foil sequences. We also predicted that the participants would improve faster on the naturalistic sequences than on the non-naturalistic sequences. To test the hypotheses regarding the incremental learning throughout the study, we ran our analyses by experimental block and exposure phase, that is, we combined blocks to determine early, intermediate, and late exposure phases, respectively.

Analysis by Experimental Block

We analyzed the data with the software R (R Core Team, 2022) using generalized linear mixed-effects models. We specified a Poisson distribution because the dependent variables (i.e., syllable and bigram recall) were count data. The models were computed using the package lmerTest 3.1-3 (Kuznetsova et al., 2017; based upon lme4 1.1-28 by Bates et al., 2015). We computed the same models with syllable recall and bigram recall as the dependent variables to test overall recall and recall of the experimental words. Models were fit with a fixed effect of sequence type using sliding contrasts (naturalistic: .5 vs. non-naturalistic: –.5, and non-naturalistic: .5 vs. foil: –.5) to examine whether

learning differed across the experimental conditions and with a fixed effect of block entered as a centered continuous variable to examine learning over the course of the experiment as well as the interaction of the two variables. We fit the maximal model supported by the data (Barr et al., 2013; Bates et al., 2018), controlling for participants and items as random intercepts, with sequence type and block as random slopes of participants (due to our within-participants design, with participants being exposed to all sequence types and blocks) but not as random slopes of items (because sequences differed between sequence types and blocks).

We checked the models for evidence of singularity in the variance-covariance matrix and for evidence of overfitting the random effects structure by conducting a principal component analysis. Models showing evidence of singularity or overfitting were simplified (for the documentation, see https://osf.io/4dsmy). To determine significance, we used an alpha level of .05. Furthermore, we have reported bootstrapped 95% confidence intervals for the beta estimates of the model predictors, based on 1,000 iterations, as well as the marginal and conditional R^2 effect sizes of the models as goodness-of-fit estimates. These R^2 values denote the proportion of the variance explained by the model both with (conditional R^2) and without (marginal R^2) controls for sources of random variance (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013).

Crucially, there was a significant main effect of sequence type at both the syllable and bigram level, with participants displaying better recall for naturalistic than non-naturalistic sequences (see Table 3), in line with our experimental hypothesis. Recall was also better for non-naturalistic sequences than for unstructured foil sequences (for a visualization of participants' syllable and bigram recall accuracy, see Figure 2 and Figure 3, respectively).

Regarding participants' performance over time, there was a main effect of block, with participants improving over the course of the experiment. Critically, there was also a significant interaction of sequence type and block, with participants' recall improving more rapidly over the course of the experiment for naturalistic sequences than for non-naturalistic sequences. However, at the bigram level, this did not meet the alpha level that we had chosen for determining significance. Participants did not improve significantly over time when recalling the non-naturalistic sequences in comparison to the unstructured foil sequences; however, participants' recall of non-naturalistic sequences numerically improved after the break at the halfway point between Blocks 6 and 7. The random-effects structure improved the model-fit in both cases.

Table 3 Summary of the linear mixed-effects models investigating the influence of sequence type and block on participants' syllable and bigram recall

Parameter	*b*	95% CI	*SE*	*t*	*p*
Syllable level					
(Intercept)	1.00	[0.87, 1.12]	0.06	15.90	< .001
Naturalistic vs. Non-naturalistic	0.10	[0.06, 0.14]	0.02	4.70	< .001
Non-naturalistic vs. Foils	0.04	[0.00, 0.07]	0.02	2.20	.03
Block	0.11	[0.07, 0.14]	0.02	6.65	< .001
Naturalistic vs. Non-naturalistic × Block	0.03	[0.01, 0.06]	0.01	2.28	.02
Non-naturalistic vs. Foils × Block	0.03	[–0.01, 0.05]	0.02	1.72	.09
Bigram level					
(Intercept)	–0.12	[–0.29, 0.04]	0.09	–1.43	.15
Naturalistic vs. Non-naturalistic	0.17	[0.10, 0.23]	0.03	4.97	< .001
Non-naturalistic vs. Foils	0.09	[0.03, 0.15]	0.03	3.02	.003
Block	0.14	[0.11, 0.18]	0.02	7.35	< .001
Naturalistic vs. Non-naturalistic × Block	0.04	[0.00, 0.09]	0.02	1.88	.06
Non-naturalistic vs. Foils × Block	0.03	[–0.03, 0.08]	0.03	1.11	.27

Note. Model fit syllable level: AIC = 10,975; BIC = 11,052; R^2_{marginal} = .079; $R^2_{\text{conditional}}$ = .091; model fit bigram level: AIC = 7,047; BIC = 7,125; R^2_{marginal} = .058; $R^2_{\text{conditional}}$ = .313.

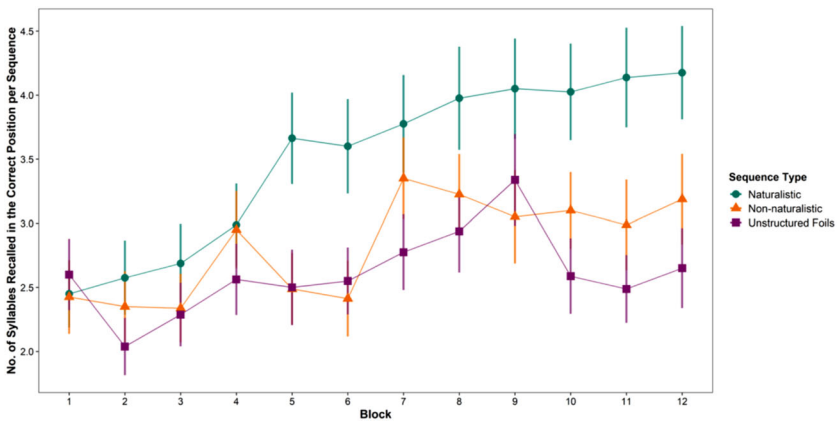

Figure 2 Mean recall of syllables (out of eight per sequence) for the three sequence types given by experimental Blocks 1–12. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. Error bars indicate ±1 standard error.

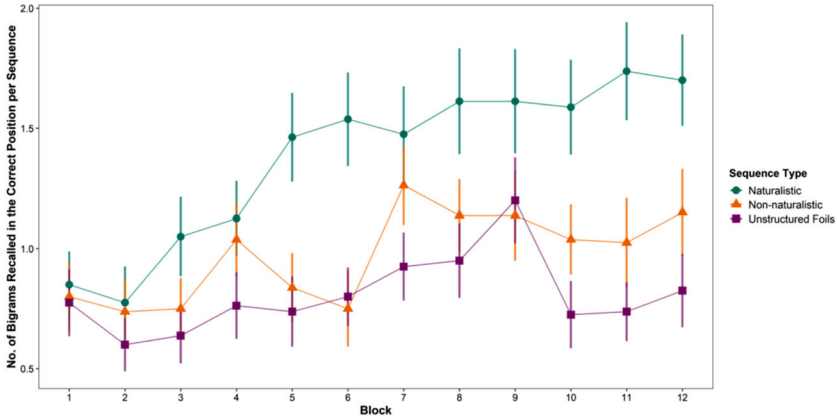

Figure 3 Mean recall of bigrams (out of four per sequence) for the three sequence types given by experimental Blocks 1–12. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. Error bars indicate ±1 standard error.

Analysis by Exposure Phase

Although the above results depicted the participants' overall improvement throughout the entire experiment, they did not reveal at which point learning began to emerge within the task. To unpack this, we divided the experiment into three phases (early exposure: Blocks 1–4; intermediate exposure: Blocks 5–8; late exposure: Blocks 9–12), testing the hypothesis that learning in the naturalistic condition would be faster than in the non-naturalistic condition. The variable exposure phase was added as a fixed effect into a new analysis instead of block. We fit the maximal model supported by the data (Barr et al., 2013; Bates et al., 2018) with sequence type (sliding contrast: naturalistic: .5 vs. non-naturalistic: –.5, and non-naturalistic: .5 vs. foil: –.5) and exposure phase (sliding contrast: early exposure: –.5 vs. intermediate exposure: .5, and intermediate exposure: –.5 vs. late exposure: .5) as well as their interaction as fixed effects, and random intercepts and slopes for participants and items, where appropriate (as described previously).

In addition to a significant main effect of sequence type, there was a main effect of exposure phase, with the participants improving between the early and intermediate exposure phase (see Table 4 for the analysis at the syllable level and Table 5 for the analysis at the bigram level; for figures illustrating the syllable and bigram recall accuracy over the three phases see the analysis folder on the project's OSF page at https://osf.io/4dsmy). The participants also improved

Table 4 Summary of the linear mixed-effects model investigating the influence of sequence type and exposure phase on participants' syllable recall

Parameter	*b*	95% CI	*SE*	*t*	*p*
(Intercept)	1.00	[0.88, 1.13]	0.06	15.95	< .001
Naturalistic vs. Non-naturalistic	0.10	[0.06, 0.14]	0.02	4.73	< .001
Non-naturalistic vs. Foils	0.03	[0.00, 0.07]	0.02	2.09	.04
Early vs. Intermediate	0.09	[0.05, 0.12]	0.02	5.04	< .001
Intermediate vs. Late	0.03	[0.00, 0.06]	0.02	1.89	.06
Naturalistic vs. Non-naturalistic × Early vs. Intermediate	0.05	[0.02, 0.09]	0.02	3.18	.001
Naturalistic vs. Non-naturalistic × Intermediate vs. Late	0.00	[–0.03, 0.04]	0.02	0.22	.83
Non-naturalistic vs. Foils × Early vs. Intermediate	0.00	[–0.03, 0.04]	0.02	0.12	.91
Non-naturalistic vs. Foils × Intermediate vs. Late	0.01	[–0.02, 0.05]	0.02	0.68	.50

Note. Model fit: AIC = 10,986; BIC = 11,135; $R^2_{\text{marginal}} = .047$; $R^2_{\text{conditional}} = .344$.

Table 5 Summary of the linear mixed-effects model investigating the influence of sequence type and exposure phase on participants' bigram recall

Parameter	*b*	95% CI	*SE*	*t*	*p*
(Intercept)	–0.12	[–0.29, 0.06]	0.09	–1.44	.15
Naturalistic vs. Non-naturalistic	0.16	[0.10, 0.23]	0.03	4.96	< .001
Non-naturalistic vs. Foils	0.09	[0.03, 0.15]	0.03	3.18	.001
Early vs. Intermediate	0.14	[0.09, 0.18]	0.02	5.79	< .001
Intermediate vs. Late	0.04	[–0.01, 0.08]	0.02	1.53	.13
Naturalistic vs. Non-naturalistic × Early vs. Intermediate	0.07	[0.02, 0.13]	0.03	2.61	.009
Naturalistic vs. Non-naturalistic × Intermediate vs. Late	0.00	[–0.05, 0.05]	0.03	–0.08	.93
Non-naturalistic vs. Foils × Early vs. Intermediate	–0.01	[–0.07, 0.06]	0.03	–0.17	.86
Non-naturalistic vs. Foils × Intermediate vs. Late	0.02	[–0.05, 0.07]	0.03	0.58	.56

Note. Model fit: AIC = 7,048; BIC = 7,143; $R^2_{\text{marginal}} = .061$; $R^2_{\text{conditional}} = .313$.

numerically between the intermediate and late exposure phase, but this did not meet the level that we had set for significance. Importantly, the interaction of sequence type and exposure phase was significant, with greater improvements on naturalistic relative to non-naturalistic sequences between the early and intermediate exposure phase. There was no difference in improvement between naturalistic and non-naturalistic sequences between the intermediate and late exposure phase. Improvement in recall of the non-naturalistic sequences did not differ from the improvement in recall of unstructured foil sequences, either between the early and intermediate exposure phase or between the intermediate and late exposure phase.

Discussion

Prior Knowledge of Syllable Co-Occurrences Facilitates Statistical Learning

SL is assumed to underlie learning across many fundamental domains of cognition, most prominently language (e.g., Christiansen & Chater, 2016; Lidz & Gagliardi, 2015; Saffran & Kirkham, 2018; Saffran et al., 1996). Although the existence of a human capacity for SL is clear, precisely how SL both depends and builds upon existing knowledge is still unclear (but see Elazar et al., 2022). Past research has shown that phonotactic probability constrains SL (Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Toro et al., 2011). In our study, we asked whether participants would draw on their prior knowledge of statistical distributions of syllables to inform their learning and processing of new linguistic input. We created two experimental languages for our native German-speaking participants; one informed by the naturally occurring TPs in German, as extracted from corpora, and another that was completely devoid of attested TPs. Breaking away from the classic format of SL paradigms that typically comprise separable training and testing phases, we presented these languages using a sequence-repetition speech-production task and tracked learning across the experiment. We hypothesized that the participants' repetitions would be more accurate and would improve more rapidly for naturalistic than for non-naturalistic sequences.

As we had predicted, recall accuracy was higher for naturalistic than non-naturalistic sequences, suggesting that the participants had drawn on their prior distributional knowledge of German to process the new experimental input. Additionally, the participants' prior experience boosted further learning of the naturalistic words in the initial stages of the experiment, increasing the recall advantage of the naturalistic sequences compared to the non-naturalistic sequences between the early and intermediate exposure phases. These findings

are consistent with the idea that learners not only track the TPs between syllables but that they also draw on this knowledge when processing subsequent input (Elazar et al., 2022; Siegelman et al., 2018), which in our study led to accelerated learning of naturalistic sequences from the beginning of the experiment. Thus, what we observed could be described as a kind of Matthew effect for SL concerning syllable transitions (Merton, 1968; for similar arguments regarding literacy, see Stanovich, 1986), where those bigrams that were attested in the participants' native language provided an advantage for future learning. This interpretation is consistent with older claims from the verbal learning literature, which has long argued that learning is a historically-contingent process that builds upon past experience (Ebbinghaus, 1885, 1913).

Overall, the results support the suggestion that participants draw upon their rich repository of existing knowledge during learning (Bertels et al., 2015; Finn & Hudson Kam, 2008; Lew-Williams et al., 2011; Lew-Williams & Saffran, 2012; Mersad & Nazzi, 2011; Onnis & Thiessen, 2013; Potter et al., 2017). An important issue concerns exactly how this existing knowledge is both represented and how it subsequently aids learning. Although many details are still to be ironed out, SL for language logically involves the discovery and registration of perceptual regularities that are then redescribed into higher level representations based on existing knowledge and generalization processes. Thus, in classic domains of enquiry like speech segmentation, TPs act as initial local cues alongside others like stress to help the listener bootstrap into the language (Cutler, 2012; Mattys & Bortfeld, 2016), after which lexical knowledge provides crucial anchors and top-down expectations about new to-be-learned material (e.g., Bortfeld et al., 2005; Lew-Williams et al., 2011; Mersad & Nazzi, 2012; for further evidence of top-down influence on the learning of adjacent dependencies, see Wang et al., 2020). We did not study segmentation per se, although we have no reason to postulate a different learning mechanism to explain our results. Accordingly, we suggest that the advantage that we observed for attested bigrams derived from this existing well-entrenched lexical knowledge providing expectations about how the input is structured, acting as local attractors through which the participants could chunk the stimuli better than when they had no, or indeed incorrectly biasing, expectations such as in the non-naturalistic condition (comparable to the bias for Spanish-like foils observed by Elazar et al., 2022).

Accordingly, we suggest that the results provide support for the idea that syllable co-occurrences are tracked and become more entrenched with each encounter (Isbilen et al., 2017, 2020; Jost & Christiansen, 2017; Siegelman et al., 2018). Such entrenchment can be seen as a form of chunking that

facilitates subsequent processing and production because participants can draw on stored chunks instead of individual syllables (e.g., Christiansen & Chater, 2016; Jones, 2012; Jones et al., 2021; Perruchet & Vinter, 1998; Robinet et al., 2011). The learning advantage seen for the sequences comprising attested TPs exemplified this further, with higher accuracy and faster learning seen for sequences that adhered to a distribution that should already have been well-entrenched within the participants due to their prior experience with German. This is in line with previous studies showing that participants drew on their long-term lexical knowledge to guide recall in short-term memory tasks (e.g., Jones & Macken, 2018; Kowialiewski & Majerus, 2018; Majerus et al., 2012; for neuroimaging evidence, see Tremblay et al., 2016). Together with Elazar et al.'s (2022) study, our study demonstrated that long-term linguistic knowledge guides future learning at the level of syllable transitions, thus complementing work on phonotactic probability (e.g., Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011).

There are potential parallels between our data and those from electroencephalogram (EEG) studies that have tracked SL across familiarization. Notably, Batterink and colleagues have demonstrated that SL is a gradual, two-staged process of chunking adjacent syllables and storing them in long-term memory (Batterink, 2020; Batterink & Paller, 2017). The properties of the EEG signal suggested that participants initially treated the speech signal as a stream of syllables. However, across familiarization participants entrained to higher levels of linguistic organization as the syllables in the stream became more familiar to them, that is, participants were able to identify that some adjacent syllables frequently co-occurred and treated them as word-like, storing these frequently co-occurring syllable combinations in long-term memory. With this in mind, one interpretation of our data is that our participants were building upon their attested knowledge of German syllabic regularities to implicitly treat naturalistic syllable pairs as word-like sooner than they did in the non-naturalistic condition, thus accounting for the difference in learning rate during the early and intermediate exposure phases of the experiment. Acquiring word-like representations of the non-naturalistic sequences compared to the foil sequences was more difficult for the participants and did not interact with the exposure phase. There are two likely reasons for this. First, the non-naturalistic sequences contained unattested TPs, and so, given the assumption that these matter, the participants were starting from the lowest of bases. Second, the non-naturalistic sequences contained the same syllables as the naturalistic sequences, which means that there could have been some interference from long-term knowledge of German.

Related to this latter point, the comparative difficulty that the participants experienced with learning the non-naturalistic sequences might have been due to the difficulty of simultaneously learning multiple languages, especially with the same syllables occurring in multiple words across sequence types. This is consistent with work by Page et al. (2013), who have shown that Hebbian learning is slower when structured sequences have item overlap (see also Antovich & Graf Estes, 2018, for evidence that bilingual but not monolingual infants can extract words from multiple experimental languages when these languages are presented interleaved). In previous studies of SL, adults learned only the first of two subsequently presented artificial languages, unless (a) there were contextual cues indicating the change between languages or (b) the exposure to the second language was either tripled or initiated before a certain level of entrenchment was reached for the words of the first language (Bulgarelli & Weiss, 2016; Gebhart et al., 2009). In our study, this level of entrenchment had presumably already been reached for the naturalistic words when the participants entered the experiment, such that participants' predisposition for (and enhanced learning of) the naturalistic words might have biased learning in favor of the naturalistic sequences at the expense of the others. Ultimately, however, there was significant evidence that the participants did learn in the non-naturalistic condition compared to the foil condition, and so acquiring multiple syllable transitions across different sequence types, even when the sequence types were drawn from the same syllable inventory, was not impossible in the context of the task.

The Serial Recall Task As a Window Into Statistical Learning

On a methodological note, this study offers an alternative behavioral method to track SL in real time, with participants' training and testing being critically intertwined. This method builds on the classic Hebbian repetition paradigm (Hebb, 1961; see also Page & Norris, 2009; Smalle et al., 2016; Szmalec et al., 2012), as well as on more recent studies that have used recall tasks to examine learning after a period of exposure to a new artificial language (Isbilen et al., 2018, 2020; Kidd et al., 2020; Majerus et al., 2004). Here, we have shown that recall tasks of this nature, in the absence of an initial exposure phase, can serve as an insightful window onto learning and may be an advantageous method for future studies of SL. Accordingly, we believe that the task can serve as a valuable addition to the toolkit of methods used to study SL. One advantage of the task that we have already discussed is the ability to use it to track learning across the course of an experiment. Another notable benefit of the repetition paradigm is that it is, in the words of Christiansen (2019), a processing-based

measure of learning. This contrasts with reflection-based measures of SL, such as traditionally used measures of SL like the 2AFC task. The difference between the two is that processing-based tasks require less meta-cognitive effort because, unlike reflection-based measures, they do not ask participants to reflect upon and choose between two possible candidate words. Although 2AFC tasks have their advantages, there are circumstances under which they are not always optimal, including when testing auditory SL in developmental populations and when the aim is to measure individual differences (see Arnon, 2019; Isbilen et al., 2020, 2022; Kidd et al., 2020). Our suggestion is that verbal repetition may be particularly useful in circumstances where researchers are interested in the course of learning or when reflection-based measures such as 2AFC do not yield reliable results.

With this in mind, one obvious question concerns exactly how verbal recall relates to other measures of SL and to the bigger question of how it relates to the mechanism underlying SL (or the multiple interacting mechanisms underlying it, see Frost et al., 2015). These questions are not mutually exclusive, and we cannot hope to provide a compelling answer to them here. What is clear is that there are many different measures of SL, going from verbal repetition to sequence reproduction (e.g., Conway et al., 2010) to 2AFC following familiarization (e.g., Saffran et al., 1996) to reaction times to structured sequences, as in the serial reaction time task (Nissen & Bullemer, 1987). It is interesting that, although all the measures capture learning of probabilistic distributions and thus are billed as measures of SL, performances on these tasks are often unrelated (e.g., Siegelman & Frost, 2015). There are likely to be many reasons for this. One obvious methodological reason is that any mode of measurement is an imperfect way of tapping a psychological concept, and so any one task will have nonoverlapping measurement error that it does not share with other tasks. More interestingly, the processes underlying SL have been argued to be complex and multi-componential (Arciuli, 2017; Frost et al., 2015), and thus different tasks may differentially implicate different components. This lack of understanding of these individual components limits the understanding of the mechanism(s) underlying SL.

What we see as the value of verbal sequence repetition is in its potential for elucidating the role of SL in language learning. Repetition has had a long history of use in the verbal learning literature beginning with Hebb (1961) and has also been used to measure linguistic proficiency. For instance, non-word repetition is highly sensitive to speakers' distributional knowledge of their language (e.g., Jones et al., 2007, 2014; Szewczyk et al., 2018), and sentence repetition reliably taps grammatical parsing procedures underlying sentence production

and comprehension (Acheson & MacDonald, 2009; Potter & Lombardi, 1990). Thus, verbal sequence repetition appears to be a relatively direct way of observing both (a) existing knowledge and, as we have shown here, (b) how that knowledge may result in different learning trajectories across time. Studying verbal repetition in a learning paradigm, as we have done in our study (see also Isbilen et al., 2020), is one way to study the dynamics of SL across time (see also Batterink, 2020; Batterink & Paller, 2017).

Limitations and Future Directions

Our results, alongside those of Elazar et al. (2022), reveal positive evidence in favor of the argument that humans identify frequently occurring linguistic units and encode them as long-term memory representations that are subsequently used for future learning. A key promise of this effect is that it captures what is assumed to be the output of SL; participants are better at learning naturalistic distributions because they have prior experience with them, distributions that they have presumably discovered via SL. However, as with most laboratory-based studies of SL, we have only tested the learning of simple statistical computations. How this scales up to the acquisition of language proper, with all of its complexities, is unclear. Domain-general processes like chunking no doubt play an important role in acquisition and in processing (e.g., Bannard & Matthews, 2008; Christiansen & Chater, 2016; Jones et al., 2021; Lieven et al., 1997). Indeed, Isbilen et al. (2022) have recently shown that adults' chunking of syllables in verbal repetition is related to their recall of highly frequent sequences of words, suggesting a partially shared basis for learning and processing across the different linguistic levels. However, it is important to be mindful of the limits of such effects as they relate to the entirety of language. In particular, because studies of SL typically limit themselves to formal aspects of language (i.e., relationships between linguistic elements devoid of meaning), how a process like SL works within the maelstrom of natural language and how it works in concert with other key learning mechanisms is still very much an open question and thus a matter for future research.

Conclusion

To conclude, in this study, we demonstrated that adult participants' prior knowledge of TPs derived from their native language forms a robust foundation upon which subsequent learning and processing occur. Our data thus lend further support to the notion that prior knowledge can have a critical impact on future learning (Bertels et al., 2015; Dal Ben et al., 2021; Ebbinghaus, 1885, 1913; Finn & Hudson Kam, 2008; Lew-Williams

et al., 2011; Lew-Williams & Saffran, 2012; Mersad & Nazzi, 2011; Onnis & Thiessen, 2013; Potter et al., 2017), providing further evidence that laboratory-based learning is shaped by the (mis/)alignment between the properties of the input and participants' prior expectations. Implementing a sequence repetition task in the absence of a familiarization phase provided a rich real-time behavioral assessment of SL (though see Batterink, 2020, and Batterink & Paller, 2017, for related online assessments using EEG). We suggest that dynamic speech-production measures may serve as a useful vehicle for further exploring the nature and time course of SL in future research.

Final revised version accepted 14 June 2022

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://osf.io/4dsmy/. All proprietary materials have been precisely identified in the manuscript.

Notes

1 Elazar et al.'s (2022) paper was published during our review process.

2 We included the following corpora from the CHILDES database (MacWhinney, 2000) in our analysis: Caroline (Von Stutterheim, 2010), Grimm (Grimm, 2006, 2007), Leo (Behrens, 2006), Manuela (Wagner, 2006), Miller (Miller, 1979), Rigol (Rigol, 2007), Stuttgart (Lintfert, 2010), TAKI (Lintfert, 2010), and Wagner (Wagner, 1974, 1985).

3 In an analysis of TPs in child-directed speech across nine languages, Saksida et al. (2017) reported a mean between-word TP of .11, compared to a mean within-word TP of .25, whereas for German, Stärk et al. (2022a) reported a mean between-word TP of .11 and a mean within-word TP of .33. Thus, our naturalistic words were, on average, more indicative of word-like units than between-word transitions.

References

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

Antovich, D. M., & Graf Estes, K. (2018). Learning across languages: Bilingual experience supports dual language statistical word segmentation. *Developmental Science*, *21*(2), 50–68. https://doi.org/10.1111/desc.12548

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), Article 20160058. https://doi.org/10.1098/rstb.2016.0058

Arnon, I. (2019). Statistical learning, implicit learning, and first language acquisition: A critical evaluation of two developmental predictions. *Topics in Cognitive Science*, *11*(3), 504–519. https://doi.org/10.1111/tops.12428

Audacity Team. (2018). *Audacity(R): Free audio editor and recorder* (Version 2.4.2) [Computer software]. https://audacityteam.org

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious mixed models*. ArXiv:1506.04967 [Stat.ME]. http://arxiv.org/abs/1506.04967

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

Batterink, L. J. (2020). Syllables in sync form a link: Neural phase-locking reflects word knowledge during language learning. *Journal of Cognitive Neuroscience*, *32*(9), 1735–1748. https://doi.org/10.1162/jocn_a_01581

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. https://doi.org/10.1016/j.cortex.2017.02.004

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and Cognitive Processes*, *21*(1–3), 2–24. https://doi.org/10.1080/01690960400001721

Bertels, J., Destrebecqz, A., & Franco, A. (2015). Interacting effects of instructions and presentation rate on visual statistical learning. *Frontiers in Psychology*, *6*, Article 1806. https://doi.org/10.3389/fpsyg.2015.01806

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Bulgarelli, F., & Weiss, D. J. (2016). Anchors aweigh: The impact of overlearning on entrenchment effects in statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1621–1631. https://doi.org/10.1037/xlm0000263

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468–481. https://doi.org/10.1111/tops.12332

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, Article e62. https://doi.org/10.1017/S0140525X1500031X

Conway, C. M., Baurnschmidt, A., Huang, S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

Dal Ben, R., de Hollanda Souza, D., & Hay, J. F. (2021). When statistics collide: The use of transitional and phonotactic probability cues to word boundaries. *Memory & Cognition*, *49*, 1300–1310. https://doi.org/10.3758/s13421-021-01163-4

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College Press (Original work published 1885).

Elazar, A., Alhama, R. G., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the "Tabula" is anything but "Rasa:" What determines performance in the auditory statistical learning task? *Cognitive Science*, *46*(2), Article e13102. https://doi.org/10.1111/cogs.13102

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*, 477–499. https://doi.org/10.1016/j.cognition.2008.04.002

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

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. https://doi.org/10.1016/j.tics.2014.12.010

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. https://doi.org/10.1111/j.1551-6709.2009.01041.x

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grimm, A. (2006). Intonational patterns and word structure in early child German. In D. Bamman, T. Magnitskaia, & C. Zaller (Eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development* (pp. 237–248). Cascadilla.

Grimm, A. (2007). *The development of early prosodic word structure in child German: Simplex words and compounds*. (Unpublished doctoral dissertation). Universität Potsdam.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37–46). Blackwell.

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

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition*, *225*, Article 105123. https://doi.org/10.1016/j.cognition.2022.105123

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 564–569). Cognitive Science Society. http://hdl.handle.net/11858/00-001M-0000-002E-3261-B

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*, *44*(7), Article e12848. https://doi.org/10.1111/cogs.12848

Johnson, E. K. (2016). Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics*, *2*(1), 391–412. https://doi.org/10.1146/annurev-linguistics-011415-040616

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2 GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

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

Jones, G., Cabiddu, F., Andrews, M., & Rowland, C. F. (2021). Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *Journal of Memory and Language*, *119*, Article 104232. https://doi.org/10.1016/j.jml.2021.104232

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: The case of nonword repetition. *Developmental Science*, *17*(2), 298–310. https://doi.org/10.1111/desc.12111

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: A computational model of the learning of new words. *Developmental Science*, *10*(6), 853–873. https://doi.org/10.1111/j.1467-7687.2007.00638.x

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory & Cognition*, *46*(2), 216–229. https://doi.org/10.3758/s13421-017-0759-3

Jost, E., & Christiansen, M. H. (2017). Statistical learning as a domain-general mechanism of entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (1st ed., pp. 227–244). Walter de Gruyter.

Jusczyk, P. W. (2002). How infants adapt speech-processing capacities to native-language structure. *Current Directions in Psychological Science*, *11*(1), 15–18. https://doi.org/10.1111/1467-8721.00159

Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E. S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology*, *200*, Article 104964. https://doi.org/10.1016/j.jecp.2020.104964

Kowialiewski, B., & Majerus, S. (2018). The non-strategic nature of linguistic long-term memory effects in verbal short-term memory. *Journal of Memory and Language*, *101*, 64–83. https://doi.org/10.1016/j.jml.2018.03.005

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition*, *118*(2), 265–273. https://doi.org/10.1016/j.cognition.2010.11.011

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247–1252. https://doi.org/10.1111/j.1467-9280.2008.02233.x

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329. https://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. https://doi.org/10.1016/j.cognition.2011.10.007

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics*, *1*(1), 333–353. https://doi.org/10.1146/annurev-linguist-030514-125236

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219. https://doi.org/10.1017/S0305000996002930

Lintfert, B. (2010). *Phonetic and phonological development of stress in German* (Unpublished doctoral dissertation). Universität Stuttgart.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum.

Majerus, S., Martinez Perez, T., & Oberauer, K. (2012). Two distinct origins of long-term learning effects in verbal short-term memory. *Journal of Memory and Language*, *66*(1), 38–51. https://doi.org/10.1016/j.jml.2011.07.006

Majerus, S., Van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, *51*(2), 297–306. https://doi.org/10.1016/j.jml.2004.05.002

Mattys, S. L., & Bortfeld, H. (2016). Speech segmentation. In M. G. Gaskell & J. Mirković (Eds.), *Speech perception and spoken word recognition* (pp. 65–85). Psychology Press.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://hrcak.srce.hr/89395

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition*, *39*(6), 1085–1093. https://doi.org/10.3758/s13421-011-0074-3

Mersad, K., & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, *8*(3), 303–315. https://doi.org/10.1080/15475441.2011.609106

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Miller, M. (1979). *The logic of language development in early childhood*. Springer-Verlag.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), Article 20170213. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Neurobehavioral Systems. (2014). *Presentation* (Version 18.0) [Computer software]. www.neurobs.com

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284. https://doi.org/10.1016/j.cognition.2012.10.008

Page, M. P. A., Cumming, N., Norris, D., McNeil, A. M., & Hitch, G. J. (2013). Repetition-spacing and item-overlap effects in the Hebb repetition task. *Journal of Memory and Language*, *69*(4), 506–526. https://doi.org/10.1016/j.jml.2013.07.001

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3737–3753. https://doi.org/10.1098/rstb.2009.0173

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263. https://doi.org/10.1006/jmla.1998.2576

Potter, C. E., Wang, T., & Saffran, J. R. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science*, *41*(S4), 913–927. https://doi.org/10.1111/cogs.12473

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.1.3) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rigol, R. (2007). German Rigol corpus [Data set]. *TalkBank*. https://doi.org/10.21415/t50s34

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*(7), 1352–1389. https://doi.org/10.1111/j.1551-6709.2011.01188.x

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

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. https://doi.org/10.1006/jmla.1996.0032

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, *20*(3), Article e12390. https://doi.org/10.1111/desc.12390

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Smalle, E. H. M., Bogaerts, L., Simonis, M., Duyck, W., Page, M. P. A., Edwards, M. G., & Szmalec, A. (2016). Can chunk size differences explain developmental changes in lexical learning? *Frontiers in Psychology*, *6*, Article 1925. https://doi.org/10.3389/fpsyg.2015.01925

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407.

Stärk, K., Kidd, E., & Frost, R. L. A. (2022a). Word segmentation cues in German child-directed speech: A corpus analysis. *Language and Speech*, *65*(1), 3–27. https://doi.org/10.1177/0023830920979016

Stärk, K., Kidd, E., & Frost, R. L. A. (2022b). *Materials, data, and results from "Close encounters of the word kind: Attested distributional information boosts statistical learning*." Open Science Framework. https://doi.org/10.17605/OSF.IO/4DSMY

Stärk, K., Kidd, E., & Frost, R. L. A. (2022c). *Materials, data, and results from "Experiment testing the validity of our stimuli*." Open Science Framework. https://doi.org/10.17605/OSF.IO/P9FCM

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition*, *179*, 23–36. https://doi.org/10.1016/j.cognition.2018.06.002

Szmalec, A., Page, M. P. A., & Duyck, W. (2012). The development of long-term lexical representations through Hebb repetition learning. *Journal of Memory and Language*, *67*(3), 342–354. https://doi.org/10.1016/j.jml.2012.07.001

Toro, J. M., Pons, F., Bion, R. A. H., & Sebastián-Gallés, N. (2011). The contribution of language-specific knowledge in the selection of statistically-coherent word candidates. *Journal of Memory and Language*, *64*(2), 171–180. https://doi.org/10.1016/j.jml.2010.11.005

Tremblay, P., Deschamps, I., Baroni, M., & Hasson, U. (2016). Neural sensitivity to syllable frequency and mutual information in speech perception and production. *NeuroImage*, *136*, 106–121. https://doi.org/10.1016/j.neuroimage.2016.05.018

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? Relations between children's cross-situational word learning, memory, and language abilities. *Journal of Memory and Language*, *93*, 217–230. https://doi.org/10.1016/j.jml.2016.10.001

Vlach, H. A., & Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cognitive Science*, *38*(4), 757–774. https://doi.org/10.1111/cogs.12092

Von Stutterheim, C. (2010). German Caroline corpus [Data set]. *TalkBank*. https://doi.org/10.21415/t5ns5s

Wagner, K. R. (1974). *Die Sprechsprache des Kindes: Teil 1. Theorie und Analyse* [The child's spoken language: Volume 1. Theory and analysis]. Präger.

Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language*, *12*(2), 475–487. https://doi.org/10.1017/S0305000900006565

Wagner, M. (2006). *First steps to communication: A pragmatic analysis*. Narr Verlag.

Wang, F. H., Zevin, J. D., Trueswell, J. C., & Mintz, T. H. (2020). Top-down grouping affects adjacent dependency learning. *Psychonomic Bulletin & Review*, *27*(5), 1052–1058. https://doi.org/10.3758/s13423-020-01759-y

Wang, T., & Saffran, J. R. (2014). Statistical learning of a tonal language: The influence of bilingualism and previous linguistic experience. *Frontiers in Psychology*, *5*, Article 953. https://doi.org/10.3389/fpsyg.2014.00953

Zettersten, M., Potter, C. E., & Saffran, J. R. (2020). Tuning in to non-adjacencies: Exposure to learnable patterns supports discovering otherwise difficult structures. *Cognition*, *202*, Article 104283. https://doi.org/10.1016/j.cognition.2020.104283

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Experiment Testing the Validity of Our Stimuli.