COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13102

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13102

When the "Tabula" is Anything but "Rasa:" What Determines Performance in the Auditory Statistical Learning Task?

Amit Elazar,[a] Raquel G. Alhama,[b] Louisa Bogaerts,[c] Noam Siegelman,[d] Cristina Baus,[e,f] Ram Frost[a,d,g]

[a]*Department of Psychology, The Hebrew University of Jerusalem*
[b]*Department of Cognitive Science & Artificial Intelligence, Tilburg University*
[c]*Department of Experimental Psychology, Ghent University*
[d]*Haskins Laboratories, New Haven, CT*
[e]*Department of Cognition, Development and Educational Psychology, University of Barcelona*
[f]*Center for Brain and Cognition, Universitat Pompeu Fabra*
[g]*BCBL, Basque Center of Cognition, Brain and Language*

Received 29 October 2020; received in revised form 9 January 2022; accepted 11 January 2022

Abstract

How does prior linguistic knowledge modulate learning in verbal auditory statistical learning (SL) tasks? Here, we address this question by assessing to what extent the frequency of syllabic co-occurrences in the learners' native language determines SL performance. We computed the frequency of co-occurrences of syllables in spoken Spanish through a transliterated corpus, and used this measure to construct two artificial familiarization streams. One stream was constructed by embedding pseudowords with high co-occurrence frequency in Spanish ("Spanish-like" condition), the other by embedding pseudowords with low co-occurrence frequency ("Spanish-unlike" condition). Native Spanish-speaking participants listened to one of the two streams, and were tested in an old/new identification task to examine their ability to discriminate the embedded pseudowords from foils. Our results show that performance in the verbal auditory SL (ASL) task was significantly influenced by the frequency of syllabic co-occurrences in Spanish: When the embedded pseudowords were more "Spanish-like," participants were better able to identify them as part of the stream. These findings demonstrate that

Correspondence should be sent to Amit Elazar, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel. E-mail: amit.elazar@mail.huji.ac.il

learners' task performance in verbal ASL tasks changes as a function of the artificial language's similarity to their native language, and highlight how linguistic prior knowledge biases the learning of regularities.

Keywords: Statistical learning; Prior knowledge; Syllable frequency; Speech segmentation

1. Introduction

The demonstration that newborns, children, and adults can parse a continuous stream of syllables solely given their co-occurrence statistics sparked intense theoretical debates in the domain of language acquisition. The seminal finding reported by Saffran and colleagues (Saffran et al., 1996b) provided evidence that an experience-based mechanism, coined as statistical learning (SL), can account for at least some aspects of language learning. The main point of this paper was that the probability of co-occurrence of syllables in an artificial input stream can be rapidly perceived and then used to extract word boundaries. In the range of many papers that followed, SL was taken to explain not only segmentation of speech (e.g., Onnis et al., 2008), but also learning phonotactic regularities (e.g., Chambers et al., 2003), detecting interwords adjacent (e.g., Endress & Mehler, 2009) and nonadjacent (e.g., Newport & Aslin, 2004) relations, including long distance dependencies (e.g., Gomez, 2002, Saffran & Kirkham, 2018).

Perhaps the most common paradigm used in SL research is the auditory statistical learning (ASL) task, which is based on the original work by Saffran et al. (1996a).[1] In this task, participants are presented with a continuous stream of syllables (typically 12–24), which co-occur in pairs, triplets, or quadruplets forming pseudowords with full or partial regularities (e.g., ba-gi, ba-gi-du, ba-gi-du-ka; see Conway & Christiansen, 2005; Gomez, 2002; Onnis et al., 2003; Peña et al., 2002 for references). The sequence of pseudowords is randomized and repeated multiple times to create a familiarization stream made of continuous syllabic elements. Following familiarization, participants are tested, usually by conveying two-alternative forced-choice (2-AFC) decisions, on their ability to differentiate between target pseudowords (i.e., syllables that consistently co-occurred together in the stream) and foils (i.e., combinations of syllables that appeared in the stream but with a lower probability). Performance in this test is taken as a proxy for an individual's ability to extract word boundaries in continuous speech (Erickson et al., 2016; Misyak & Christiansen, 2012; Siegelman & Frost, 2015) given differences in statistical co-occurrences.

Since novel pseudowords—and not real words—are the target of learning, the underlying implicit supposition of this experimental paradigm is that participants' linguistic knowledge is irrelevant for performance. Thus, for any given individual, learning outcomes would be determined by the statistical structure of the input stream (i.e., the elements in the stream and their distributional properties). For example, syllables which co-occur systematically (with a probability of $p = 1.0$) would be easier to learn than quasiregular patterns (syllabic elements which co-occur with a probability of $p = .8$), and four embedded pseudowords would be easier to learn than eight. Note that this assumption regarding performance in the ASL task is

independent of the highly debated question whether the assimilation of the input's statistics occurs via tracking of transitional probabilities (e.g., Endress & Mehler, 2009), or via extraction of chunks (e.g., Perruchet & Vinter, 1998; Perruchet et al., 2014; and see Siegelman et al., 2019 for review and computational evidence). Importantly, independent of the assumed underlying learning mechanism, learning scores in the task are taken to reflect the learner's acquired sensitivity to the regularities embedded in the stream.

Siegelman et al. (2018) labeled this supposition the "tabula rasa" assumption, and provided arguments why it is violated when it comes to linguistic material. In a series of experiments, they demonstrated that, in contrast to learning patterns of abstract shapes in the visual modality or in nonverbal auditory material, learning of one pseudoword in the auditory stream does not predict the learning of another. This is reflected in low internal consistency of auditory verbal SL tasks. The significant variance in performance across different pseudowords in verbal ASL tasks does not coincide with the tabula rasa assumption: If learning is exclusively driven by the co-occurrence statistics of the elements, all pseudowords in the stream should be learned with an equal probability, and thereby the task should display high internal consistency (see Alhama & Zuidema, 2016, 2017 for related claim). Why then are some pseudowords learned better than others in the ASL task?

Siegelman et al. (2018) argued that the answer to this question lies, at least partially, in participants' prior knowledge of the statistical regularities in their native language. But what precisely is this "prior" knowledge? Which linguistic information modulates performance in the ASL task? Some answers come from studies showing that the phonotactics of participants' native language influence ASL performance, as participants struggle to learn pseudowords that are phonotactically illegal in their language (Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011). Similarly, developing syntactic knowledge about one's native language has been shown to influence the learning of sequential regularities as early as 13 months (Thiessen et al., 2019; see also Onnis & Thiessen, 2013 for similar results in adults). Further preliminary insights are provided in the Siegelman et al. study, Experiment 4: This experiment revealed that the perceived similarity (or dissimilarity) between the pseudowords used in the experiments and the participants' native language explained part of the variance in performance in the auditory verbal SL task. But what precisely does this "similarity" refer to? Subjective ratings of perceived similarity may be driven by a wide range of factors ranging from low-level phonetics to high-level morphology (Dijkstra et al., 2010; Robinson, 2005).

The present study targets the question regarding the nature of prior knowledge directly. Concretely, our study examined the hypothesis that the frequency of co-occurrence of syllabic units in participants' native language impacts performance in the auditory verbal SL task. Specifically, we hypothesized that native speakers of a language develop expectations regarding the co-occurrences of different syllables in their language, which would inevitably facilitate or hinder learning of an artificial language.

1.1. The present study

The goal of the current study was to examine how learning the structure embedded in the artificial stream is affected by the prior knowledge of syllable co-occurrences in the learners'

native language. We experimentally manipulated the statistical co-occurrence of syllabic elements per participants' native language when building the artificial language used in the ASL task. Most verbal ASL tasks use syllabic units (typically consonants-vowel [CV] elements), combining them without explicit consideration of their distributional properties in the participants' native language. Exemplifying this issue, a recent review found that about a quarter of the studies using the standard ASL task used the exact same trisyllabic pseudowords as Saffran and colleagues (1996a) (Frost et al., 2019), independent of the native language of the studied sample (e.g., McNealy et al., 2006; Toro et al., 2005). In our study, we set out to carefully consider the interaction between the statistical properties of the artificial stream and participants' native language, focusing on Spanish as the target language. We constructed continuous auditory streams of trisyllabic pseudowords, while manipulating the probability of syllabic co-occurrence in Spanish.[2] Our goal was to examine whether and to what extent performance of native Spanish speakers in extracting the embedded pseudowords from a continuous auditory stream of syllables is determined by the likelihood of syllabic co-occurrences in their native language. To preview our findings, our results show that a significant part of the variance in ASL performance is determined by the frequency of syllabic co-occurrences in the learners' native language.

2. Method

2.1. Participants

One hundred and eighty-seven subjects (64 males) from two sites in Spain participated in the experiment: 85 students from the University of the Basque Country (UPV/EHU), Donostia-San Sebastián (Gipuzkoa), and 102 students of University Pompeu Fabra, Barcelona (referred to as site 1 and site 2, respectively). All participants reported Spanish as their dominant language, as confirmed by a self-assessed language proficiency questionnaire. Note that the decision to conduct the experiment in two regions of Spain was made in an attempt to increase the study's external validity, because different provinces differ in some phonetic characteristics of the spoken language or the relative prevalence of some spoken words. We return to this design feature and its potential impact on our findings in the General discussion.

2.2. Design and procedure

We created two artificial languages: one consisting of pseudowords with high co-occurrence syllable frequency in Spanish ("Spanish-like" condition) and another consisting of the same syllables, recombined to form pseudowords with low co-occurrence frequency in Spanish ("Spanish-unlike" condition; see "*Stimuli construction*" below). The syllable set included 24 CV syllables that were recorded in isolation by a Spanish native speaker. Syllables had a mean duration of 420 ms (range: 300–500 ms) and were organized into eight pseudowords in each condition. The eight pseudowords were randomized to create a 7-min familiarization stream, which contained 48 repetitions of each pseudoword. Importantly, the stream was continuous as it did not contain breaks or any other acoustic markers between the pseudowords. The only constraint in the randomization order was that the same pseudoword

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 familiarization stream was then played to participants via earphones. Following familiarization, a 20-item old/new decision task (Forest et al., 2020; Franco et al., 2011) began, with the eight pseudowords serving as targets (i.e., "old"—appearing in the stream) and 12 additional foils (i.e., "new"—syllabic combinations that did not appear in the stream). Six of those foils were unique to each condition, and six were mid-range filler-foils shared between the two conditions. During the test, each item was played in isolation and participants were instructed to judge whether it was part of the language they heard by pressing 1 if the item appeared in the stream and 2 if it did not. After conveying their response, participants were asked to rate their confidence in their answer on a 7-point Likert scale, ranging from 1 (not sure at all) to 7 (very sure) (see Alhama et al., 2015 for a similar procedure).

2.3. Stimuli construction

Our goal was to create a set of pseudowords of the form CVCVCV, such that the co-occurrences of syllables within a triplet were either frequent (in the "Spanish-like," high-frequency condition) or infrequent ("Spanish-unlike," low-frequency condition) in Spanish. To do so, we extracted CV-syllables' frequencies from Corpus Oral de Referencia del Español Contemporáneo (CORLEC, Marcos Marín, 1992), a database of transliterated spoken Spanish. The corpus contains over 1.1 million words from recordings in different domains (administrative, conversational, educational, etc.). As a first step, we converted the transcribed corpus from orthographic to pseudo-phonetic representations. For instance, in Spanish, "v" and "b" represent the same phoneme, so all "v" letters were converted into "b." We then generated a list of all possible triplet combinations, using most of Spanish consonants and vowels. A few consonants were excluded to avoid confounds due to regional variations, ambiguity, or very low frequency (see Appendix A for details). Consonants and vowels were never repeated within a triplet. Subsequently, we computed the frequency of syllable bigrams based on their appearance in the corpus: Given a triplet of the form ABC (where A, B, and C are CV syllables), we considered the summed frequency of the bigrams AB and BC in the corpus, and selected triplets that fall within the frequency thresholds of the high- ($>$400 appearances) and low-frequency ($<$200 appearances) conditions (see Appendix B for full description).

The selection of the pseudowords to serve as targets in the two conditions required the satisfaction of multiple constraints that are typically applied in ASL tasks: (1) AB, BC, and ABC cannot be words in Spanish; (2) bigrams cannot repeat across pseudowords (AB or BC of one triplet cannot be used in another); (3) there cannot be any phoneme repetitions in a triplet; (4) the pool of syllables used to construct the triplets has to be identical across conditions; and (5) the syllables should always appear in the same position in a triplet across conditions (Table 1). For example, *ni-be-mo* is a possible candidate for the "Spanish-like" condition, since both *ni-be* and *be-mo* are frequent bigrams in Spanish but they are not words in that language, and neither is *nibemo*. Similarly, *se-li-mo* is a possible candidate for the "Spanish-unlike" condition, because *se-li* and *li-mo* are infrequent bigrams in Spanish.

Table 1
Targets and foils in the two experimental conditions

High frequency		Mid frequency	Low frequency	
Target	Foil	Foil	Target	Foil
famigo	*bemosi*	*gosima*	*falume*	*dekuso*
nibemo	*demasu*	*gumeda*	*kusiga*	*limasu*
poside	*mesolu*	*fanibe*	*mabego*	*mibega*
seguna	*migona*	*lugaje*	*nilade*	*nimola*
kulame	*nikula*	*modeku*	*poguda*	*posena*
soluga	*posegu*	*senali*	*selimo*	*sigume*
felida			*sojena*	
majesu			*femisu*	

Foils for the test phase of each condition were created by shuffling the syllables of the pseudowords within each frequency class while keeping the same frequency level.[3] This was done so that participants' decisions in the test would not be based on superficial judgments of frequency of syllabic segments. We did not constrain the ordinal position of syllables in foils to match the original position of a given syllable within a pseudoword (i.e., first, second, and third positions), since it violated the frequency thresholds chosen. For example, a foil in the high-frequency condition was *be-mo-si*, in which both *be-mo* and *mo-si* are frequent bigrams in Spanish.

In addition to high- and low-frequency pseudowords and high- and low-frequency foils (unique to each condition), we created a set of mid-frequency filler-foils to be part of both test conditions. These shared filler-foils were constructed using the same syllables of the two artificial languages, but had a mid-frequency range (in-between the Spanish-like and the Spanish-unlike stimuli, summed frequency of 300–400 appearances; see Appendix B for details). The goal of adding identical mid-range filler-foils in each of the tests was to have an anchor for performance comparison across conditions. We reasoned that if subjects base their decisions in the test also on the knowledge they have regarding the statistical regularities of syllable occurrences in their language (in addition to the co-occurrence statistics in the familiarization stream they were just exposed to), they would accept Spanish-like foils more than Spanish-unlike foils. This could result in underestimating the learning of the novel pseudowords in the Spanish-like condition, which is measured by contrasting performance on foils versus targets in each condition. Since we predicted performance for the filler-foils would be comparable across conditions, it held the promise of providing a baseline against which we could assess performance for the other foil types as well as the targets.

Overall, 20 stimuli (i.e., targets and foils) were used in the test phase of each of the two conditions: eight targets and six foils unique to each condition, in addition to six shared mid-range filler-foils. This was the maximal number of stimuli that could have been created, given the problem of multiple constraints satisfaction: We implemented a form of Backtracking Search,[4] a recursive algorithm that explores the possible triplet space in random directions (Marques-Silva & Sakallah, 1999). The process selects the next random triplet that meets

all the requirements until either all triplets are selected, or no other triplet, which does not violate the constraints, is left. If the latter happens, the algorithm "backtracks" one step and selects another triplet. If backtracking one step is not enough, the algorithm goes back as many steps as necessary while efficiently searching the space, without repeating search paths. In our implementation, we did not search for the whole set of triplets on one go, but in a pipeline: We first searched and selected a list of high-frequency triplets, then used the constraints derived from the list created to search the low-frequency triplets, only then continuing to selecting foils that would be part of the test phase. This allowed us to manually inspect the selected set of triplets before generating a new set. Targets, unique foils, and shared filler-foils are presented in Table 1; the code for the stimuli selection can be found at https://github.com/rgalhama/priork/.

2.4. Old/new task

Most SL experiments use as a learning measure a 2-AFC recognition task, where participants have to decide which of two presented stimuli appeared in the familiarization stream. This decision process requires participants to compare two candidates, knowing that only one of them appeared in the stream. This inevitably leads to selection strategies, often based on elimination rather than recognition (see, e.g., Alhama, 2017; Alhama et al., 2015). Additionally, items are typically repeated in the task several times in different combinations, which may result in further learning (or interference) during the test phase itself (see Siegelman et al., 2017 for discussion).

To avoid such biases and confounds, in the present study we implemented a familiarity judgment task (Buchner, 1994; Forest et al., 2020; Gomez, 1997; Servan-Schreiber & Johnson, 1990), where in each trial participants were presented with one trisyllabic unit, a target or a foil, and were required to decide whether it has appeared in the stream ("old"), or whether it is a new word ("new"). Following each decision, participants were asked to rate their certainty regarding their decision.

3. Predictions and analyses

We used four dependent measures to test for differences between Spanish-like and Spanish-unlike conditions in the old/new task: overall task performance, signal detection theory (SDT) measures of sensitivity and bias, and certainty ratings.

3.1. Task performance

Task performance was defined as recognition accuracy in the old/new task. Thus, both accepting targets and rejecting foils constituted correct responses, whereas rejecting targets and accepting foils constituted incorrect responses. In our main analysis below, we tested the interaction between language similarity (Spanish-like or Spanish-unlike) and item type (target/foil), on data from trials presenting targets and unique foils. Specifically, we hypothesized that the recognition of targets ("old") and detection of foils ("new") will be driven by their resemblance to words in Spanish, so that subjects would more easily accept Spanish-

like targets than Spanish-unlike targets, but will more easily reject Spanish-unlike foils than Spanish-like foils. We thus predicted that targets that are composed of syllabic elements that frequently co-occur in Spanish will be better recognized in the test phase, and foils that represent infrequent syllabic combinations will be more easily recognized as "new," and thus will be correctly rejected.

3.2. Signal detection theory measures: Response bias and sensitivity

To assess participants' ability to differentiate "old" from "new" trials, we used two measures of SDT: sensitivity (*d'*) and response bias (*c*) (Lynn & Barrett, 2014; Macmillan & Creelman, 2004). The advantage of SDT is that it allows combining responses to targets (the "signal") and foils (the "noise") to create a unified measure of sensitivity to the signal in the noise. The observer's ability to differentiate between targets and foils is reflected in the distance between the mean of the signal distribution and the mean of the noise distribution. The sensitivity measure is thus calculated as the difference between the z-transformed probabilities of hits and false alarms:

$$d' = z\,(Hit\ Rate) - z\,(FA\ rate)$$

A value of $d' = 0$ indicates inability to distinguish signal from noise, and $d' > 0$ indicates a greater ability to distinguish signal from noise. A $d' = 2$, for example, indicates that the distance between the means is twice as large as the standard deviations of the two distributions (Stanislaw & Todorov, 1999).

The added value of using SDT is that, aside of measuring sensitivity, it also monitors response biases in decisions (*C*): The tendency to use one of the two responses ("Yes," "No") more frequently than the other. *C* is the distance between the decision criterion and the neutral point, which is located at the intersection of the signal and noise distributions, where neither response is favored.

$$c = \frac{z\,(Hit\ Rate) + z\,(FA\ rate)}{-2}$$

The measure of bias allows us to examine whether syllabic sequences that are frequent in Spanish incur a "Yes" response regardless of whether they are targets or foils. If this will be the case, it will result in a larger response bias in the Spanish-like condition, where both targets and foils are similar to Spanish.

4. Results

4.1. Task performance

4.1.1. Spanish-like and Spanish-unlike targets and foils

The task performance results in trials with targets and unique foils are shown in Fig. 1. As can be seen, subjects were able to correctly accept targets well above chance, suggesting they learned the pseudowords in both the Spanish-like (Mean = 78.7%, $SD = 16.7\%$, $t_{(93)} = 16.6$, $p < .001$) and Spanish-unlike (Mean = 73.1%, $SD = 15.5\%$, $t_{(92)} = 14.36$, $p < .001$)

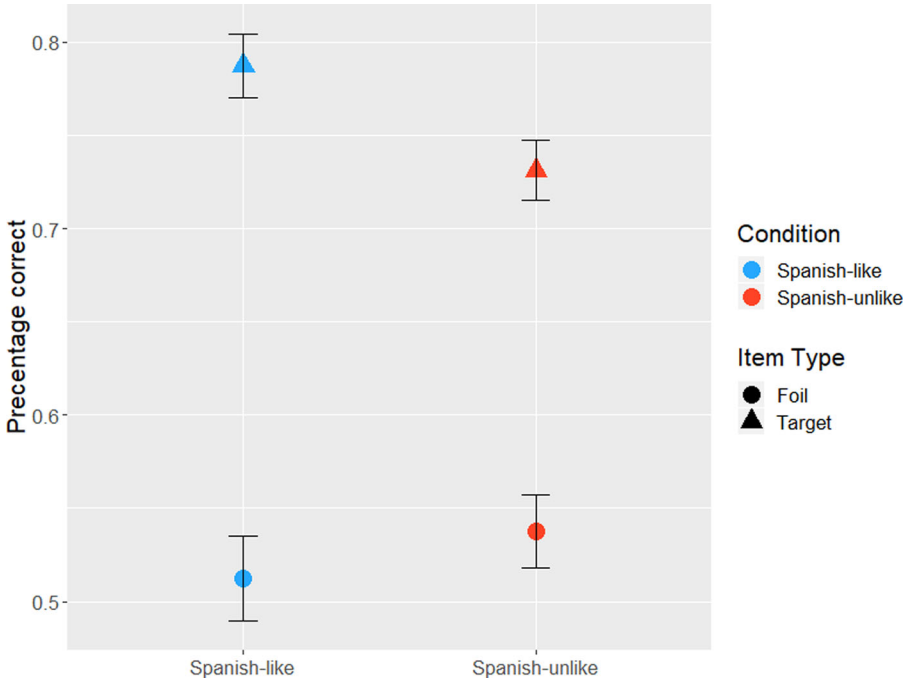

Fig 1. Task performance by condition (Spanish-like vs. Spanish-unlike) and trial type (targets vs. unique foils).

conditions. Performance for the foils in the Spanish-like condition did not significantly differ from chance (Mean = 51.2%, SD = 22.1%, $t_{(93)} = 0.523$, $p = .6$), while in the Spanish-unlike condition, it was marginally significant (Mean = 53.7%, SD = 18.8%, $t_{(92)} = 1.89$, $p = .06$). Note that a very similar pattern was observed in both sites (Fig. 2). Further analyses are, therefore, conducted on the merged dataset, with site (Barcelona or San Sebastián) as a control variable.

To examine the interaction effect of language similarity (Spanish-like/Spanish-unlike condition), item type (target/foil) and site (Barcelona/San Sebastián), as well as all main effects on performance in the task, we conducted a mixed design ANOVA,[5] with condition and site as between-subject factors and item type as a within-subject factor.

A main effect for item type ($F(1,183) = 123.84$, $p < .001$) was found: Subjects were more likely to recognize targets as "old" than foils. Importantly, the hypothesized interaction effect between language similarity and item type was, indeed, significant ($F(1,183) = 3.95$, $p = .048$), reflecting a bigger difference in performance between targets and foils in the Spanish-like versus Spanish-unlike condition (Fig. 1). No interaction was found between site, condition, and item type ($F(1,183) = 0.154$, $p = .69$), meaning the pattern was similar across sites. Main effect of site ($F(1,183) = 8.09$, $p = .005$) and its interaction with item type ($F(1,183) = 3.92$, $p = .049$) were also significant: The difference between targets and foils was smaller in Barcelona than in San Sebastián (i.e., participants had better learning overall in San Sebastián). All other effects were not significant ($p > .35$).

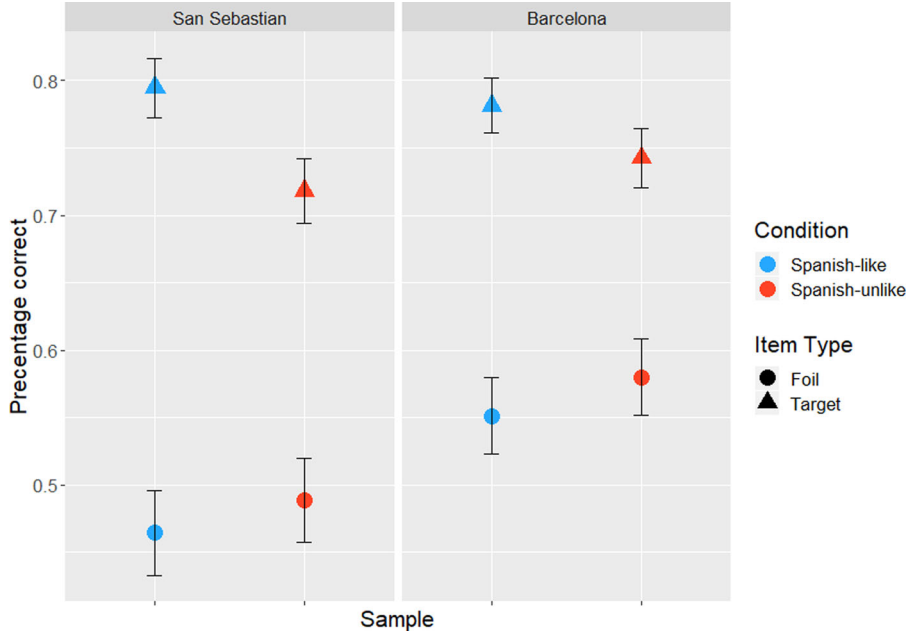

Fig 2. Task performance by condition (Spanish-like vs. Spanish-unlike) and trial type (targets vs. unique foils) in the two sites.

4.1.2. Shared filler-foils

We employed the shared filler-foils as anchor, as we originally predicted that they will be judged similarly across conditions, producing a similar error rate. However, examining the mean performance of the shared filler-foils revealed a surprisingly better performance in the Spanish-like condition ($M = 64.5\%$, $SD = 22.5\%$) than in the Spanish-unlike condition ($M = 55.5\%$, $SD = 20\%$). This difference was significant ($t_{(183)} = 2.88$, $p = .004$), indicating that subjects were better in rejecting these foils in the Spanish-like condition. This is an intriguing finding, demonstrating yet again how judgment of verbal stimuli is impacted by the linguistic environment in which these stimuli are embedded. We return to this finding in the General discussion.

4.1.3. Signal detection theory measures

We calculated sensitivity and bias measures, as described in the Predictions and analyses section. Results are presented in Table 2.

4.2. Sensitivity

In both conditions, *d'* significantly differed from zero, suggesting that subjects were able to differentiate signal from noise (Spanish-like: $t_{(93)} = 9.33$, $p < .001$; Spanish-unlike: $t_{(92)} = 9.04$, $p < .001$). However, importantly, per our hypothesis, discrimination between signal and noise was better when stimuli were similar to Spanish (Spanish-like $d' = 1.25$,

Table 2
Rates of hits and misses, false alarms, and correct rejections in the two experimental conditions, along with measures of sensitivity (*d'*) and response bias (*c*)

		Target	Foil	Sensitivity *(d')*	Response bias *(c)*
Spanish-like	Yes	Hit = 0.787	FA = 0.487	1.25	–0.54
	No	Miss = 0.213	CR = 0.513		
Spanish-unlike	Yes	Hit = 0.731	FA = 0.462	0.9	–0.36
	No	Miss = 0.269	CR = 0.538		

Table 3
Contrasts, means, and standard deviations of the certainty ratings

Item type	Spanish-like condition	Spanish-unlike condition	$t(df = 366)$	p
Targets	5.47 ($SD = 0.74$)	5.25 ($SD = 0.77$)	1.83	.06
Unique foils	4.66 ($SD = 0.86$)	4.59 ($SD = 0.80$)	0.58	.56
Filler foils	4.69 ($SD = 0.85$)	4.56 ($SD = 0.81$)	1.06	.29
Total	4.94 ($SD = 0.90$)	4.80 ($SD = 0.85$)	2.14	.14

Spanish-unlike $d' = 0.9$, $t_{(185)} = 2.06$, $p = .04$). These results are consistent with the task performance analysis above, which showed a greater difference between targets and foils in the Spanish-like condition than in the Spanish-unlike condition.

4.3. Response bias

In both conditions, bias scores were significantly lower than zero, reflecting an overall bias to say "yes" in both conditions (Spanish-like $c = –0.54$, $t_{(93)} = 7.24$, $p < .001$; Spanish-unlike $c = –0.36$, $t_{(92)} = 6.41$, $p < .001$). The "yes" response bias, however, was marginally larger in the Spanish-like condition ($t_{(185)} = 1.84$, $p = .06$), reflecting a tendency to identify both targets and foils of this condition as previously heard.

4.4. Certainty ratings

Certainty ratings regarding decisions are presented in Table 3. Marginally higher certainty was reported for targets in the Spanish-like condition than the Spanish-unlike condition ($p = .06$). The differences for unique foils and shared filler-foils did not reach significance.

5. General discussion

In this study, we aimed to explore how prior linguistic knowledge determines learning performance in the ASL task. Since performance in the task is taken to reflect sensitivity to co-occurrence statistics, we chose to focus on the frequency of co-occurrence of adjacent syllables in participants' native language. Our hypothesis was that this feature would impact the

probability of learning the pseudowords embedded in the auditory stream. We, therefore, constructed trisyllabic sequences that have high or low frequency of co-occurrence in a database of transliterated spoken Spanish, testing Spanish participants.

To assess learning outcomes, we considered participants' overall performance, as well as measures of sensitivity and bias from SDT. Our findings show that frequency of syllabic co-occurrences significantly impacted learning in the ASL task. Participants' recognition of artificial target words was better if their bi-syllabic sequences were highly frequent in Spanish than if they were infrequent, and so was their sensitivity in discriminating targets from foils. However, the response bias measure revealed that, additionally, frequent syllabic co-occurrences led to an overall increased tendency to judge pseudowords as "old."

Our experiment was conducted in two different sites in Spain. Whereas all our participants were dominant in Spanish, it is worth pointing out that these regions coexist with other languages, mainly Basque or Catalan. This factor could have potentially affected the extent of exposure to sublexical Spanish syllabic units, thus interfering with our Spanish-likeness manipulation. Nevertheless, the pattern of results revealed a clear effect of "Spanish-likeness," in spite of the additional language spoken in these parts of Spain, which was very similar across sites. Our findings thus provide a theoretical account as to why pseudowords embedded in a continuous auditory stream display low internal consistency, so that some syllabic sequences are learned better than others (Siegelman et al., 2018). Although factors such as first-language phonotactics (Finn & Hudson Kam, 2008) and variance in word length (Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012) have been reported to interfere with detection of word boundaries, frequency of syllabic co-occurrence seems to be a critical variable that determines learning in the task. While additional research should clarify how perversive this effect is across stimuli, it seems evident that the implicit assumption that performance in the ASL task is solely determined by differences in co-occurrence statistics of the syllabic units should be revised.

The present findings have, therefore, important methodological implications. Performance of participants in the task is often compared across groups (e.g., children with SLI vs. typically developing children, Evans et al., 2009, and see Saffran & Kirkham, 2018 and Bogaerts et al., 2021 for reviews), or is taken to predict a given cognitive function (see Siegelman et al., 2017 for discussion). This leads to a state of affairs where individual scores or group means in the task would be determined to a large extent by the specific syllabic sequences employed in the experiment. Scores would, therefore, increase or decrease simply by selecting a specific set of artificial words. Frequency of syllabic co-occurrences constitutes an important prior knowledge feature that participants bring with them into the experiment when they perform the task, it may hinder or improve their performance according to the specific selected stimuli for the study.

In the present study, we opted for using an old/new test rather than the common 2-AFC test in which participants are required to compare targets and foils (see Alhama et al., 2015 for discussion). There are two important advantages in this choice. First, it lends itself to SDT measures, which simultaneously tap participants' sensitivity in discriminating targets and foils, and also their response biases. Second, it allows a direct and independent measure of target word learning, which is not based on strategically eliminating foils. In other

words, 2-AFC tests cannot differentiate between decisions based on true identification of target words, and decisions based on rejecting what appears to be the strange foil. Our finding that the shared filler-foils incurred different rejection rates if they were part of the Spanish-like or Spanish-unlike condition demonstrates the potential problems in assessing learning when responses can be based on rejecting foils and not recognizing the targets. In the old/new task, each target word is judged on its own, allowing us to pinpoint which items were learned and which were not. We wish to emphasize, however, that the prior knowledge effects we documented might be generalized to the more standard 2-AFC test format. That is, we would predict performance in 2-AFC tasks to be impacted by the similarities or dissimilarities of both targets and foils to the co-occurrence statistics of participants' native language. Future research can further clarify.

To conclude, our findings demonstrate how prior linguistic knowledge regarding syllabic co-occurrences in a language significantly determines performance in auditory verbal SL tasks. The Tabula is, indeed, anything but rasa when linguistic stimuli are part of an input stream. Assessing extent of individual ability in SL tasks that involve verbal material requires, therefore, significant scrutiny. Furthermore, theories of assimilation of verbal auditory streams should consider not only the statistics of the novel input, but also the prior knowledge brought upon by participants to the learning scenario. Quantifying the range of statistical regularities that characterize different languages presents itself as an important challenge for future research. The current study directly speaks to the great promise of a corpus approach.

Acknowledgment

We thank Larraitz López for her help in running the Experiment in the BCBL.

Funding

This paper was supported by the ERC Advanced Grant (project 692502, L2STAT), awarded to Ram Frost under the Horizon 2020 research and innovation program.

Conflicts of interest

The authors have no conflicts to disclose.

Notes

1 The seminal paper by Saffran et al. (1996a) reported an experiment conducted with infants and so no explicit decisions was involved in the offline test. We refer here to the parallel design used extensively with older children and adults (e.g., Saffran et al., 1996b).

2 We selected Spanish as a target language because of its simple vowel system (five vowels), allowing a simple construction of CV segments, and its transparent orthography, which allows for a straightforward conversion from transcribed orthographic forms to phonology (see below).
3 The important advantage in constructing the foils by shuffling syllables rather than using part-words is that it enabled us to assure that TP between every two syllables in a foil is 0, and not greater. This was important given our focus on extent of stimuli's similarity to Spanish.
4 The constraints listed above pose a great limitation on finding a set of triplets that meets all the requirements at the same time, even though the potential triplet space is very large. For example, if we choose the high-frequency stimulus famigo, we need to find three triplets for the low-frequency class such that (1) the first syllable of one is fa, the second syllable of another is mi, and the third of the latter is go; (2) these syllables were not repeated; (3) there were no letter repetitions within a triplet; (4) the syllable bigrams did not overlap with the words in the high-frequency class, and these bigrams' frequencies were within the thresholds of the low-frequency class. Given all these multiple constraints, the options were reduced drastically after every triplet selection. In addition, syllables used for these three triplets would not be available to generate the rest of the triplets in this class, so every step becomes increasingly more constrained. We, therefore, needed to use a method that allows us to perform a search in a big combinatorial space, from which there were very few solutions.
5 We initially performed a logistic mixed-effect model that did not converge with any of the random effects. For this reason, we decided to conduct an ANOVA instead.

References

Alhama, R. G., Scha, R., & Zuidema, W. (2015). How should we evaluate models of segmentation in artificial grammar learning? In *Proceedings of 13th International Conference on Cognitive Modeling* (pp. 172–173).

Alhama, R. G., & Zuidema, W. (2016). Generalization in artificial language learning: Modelling the propensity to generalize. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning* (pp. 64–72).

Alhama, R. G. (2017). Computational modelling of artificial language learning: Retention, recognition & recurrence. ILLC dissertation series DS-2017-08. Institute for Logic, Language and Computation, University of Amsterdam.

Alhama, R. G., & Zuidema, W. (2017). Segmentation as retention and recognition: The R&R model. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1531–1536).

Bogaerts, L., Siegelman, N., & Frost, R. (2021). Statistical learning and language impairments: Toward more precise theoretical accounts. *Perspectives on Psychological Science*, *16*(2), 319–337.

Buchner, A. (1994). Indirect effects of synthetic grammar learning in an identification task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 550–566.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

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.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, *62*(3), 284–301.

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*(3), 351–367.

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.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*, 321–335.

Finn, A. S., & Hudson Kam, H. C. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*, 477–499.

Franco, A., Cleeremans, A., & Destrebecqz, A. (2011). Statistical learning of two artificial languages presented successively: How conscious? *Frontiers in Psychology*, *2*, 229.

Forest, T. A., Finn, A. S., & Schlichting, M. L. (2020). What is represented in memory after statistical learning? *Cognitive SciencesProceedings of the 42nd Annual Conference of the Cognitive Science Society*.1882–1888.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128.

Gomez, R. L. (1997). Transfer and complexity in artificial grammar learning. *Cognitive Psychology*, *33*(2), 154–207.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345.

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.

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological Science*, *25*(9), 1663–1673.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Marcos Marín, F. (1992). Corpus oral de referencia de la lengua española contemporánea (CORLEC).

Marques-Silva, J. P., & Sakallah, K. A. (1999). GRASP: A search algorithm for propositional satisfiability. *IEEE Transactions on Computers*, *48*(5), 506–521.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience*, *26*(29), 7629–7639.

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition*, *39*(6), 1085–1093.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302–331.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Onnis, L., Christiansen, M. H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 25, No. 25).

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284.

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, *109*(3), 423–430.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604–607.

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta Psychologica*, *149*, 1–8.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

Robinson, P. (2005). Rules and similarity processes in artificial grammar and natural second language learning: What is the "default"? *Behavioral and Brain Sciences*, *28*(1), 32–33.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.
Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181–203.
Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 592–608.
Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120.
Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213.
Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432.
Siegelman, N., Bogaerts, L., Armstrong, B. C., & Frost, R. (2019). What exactly is learned in visual statistical learning? Insights from Bayesian modeling. *Cognition*, *192*, 104002.
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.
Thiessen, E. D., Onnis, L., Hong, S. J., & Lee, K. S. (2019). Early developing syntactic knowledge influences sequential statistical learning in infancy. *Journal of Experimental Child Psychology*, *177*, 211–221.
Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34.

Appendix A

Grapheme to pseudo-phoneme substitution rules:
‘qu’: ‘k’,
‘ca’: ‘ka’,
‘co’: ‘ko’,
‘cu’: ‘ku’,
‘ge’: ‘je’,
‘gi’: ‘ji’,
‘ga’: ‘Ga’,
‘gue’: ‘Ge’,
‘gui’: ‘Gi’,
‘go’: ‘Go’,
‘gu’: ‘Gu’,
‘v’ : ‘b’,
‘ch’: ‘X’,
‘ll’: ‘L’,
‘ce’: ‘ze’,
‘ci’: ‘zi’,
‘ r’: ‘R’, #r in initial position
‘rr’: ‘R’,
‘lr’: ‘lR’,
‘sr’: ‘sR’,
‘nr’: ‘nR’

Phonemes that were not used when generating triplets:
ll, r—removed due to several optional pronunciations.
w, x, y, ñ—removed due to rarity.

Appendix B

Triplet	Condition	Type	Freq(A)	Freq(B)	Freq(C)	Freq(AB)	Freq(BC)	Freq(AB) + Freq(BC)
famigo	High	Target	2995	15,524	9502	304	377	681
nibemo	High	Target	12,697	19,670	25,129	383	417	800
poside	High	Target	28,525	30,402	77,128	1317	628	1945
seguna	High	Target	35,032	7592	28,256	1463	1024	2487
kulame	High	Target	14,158	50,441	27,117	628	653	1281
soluga	High	Target	20,090	5496	9416	543	364	907
felida	High	Target	4304	12,807	25,026	231	1370	1601
majesu	High	Target	27,906	7953	8298	200	281	481
bemosi	High	Foil	19,670	25,129	30,402	417	12	429
demasu	High	Foil	77,128	27,906	8298	1247	2	1249
mesolu	High	Foil	27,117	20,090	5496	8	543	551
migona	High	Foil	15,524	9502	28,256	377	42	419
nikula	High	Foil	12,697	14,158	50,441	3	628	631
posegu	High	Foil	28,525	35,032	7592	42	1463	1505
falume	Low	Target	2995	5496	27,117	0	21	21
femisu	Low	Target	4304	15,524	8298	14	2	16
kusiga	Low	Target	14,158	30,402	9416	45	85	130
mabego	Low	Target	27,906	19,670	9502	38	16	54
nilade	Low	Target	12,697	50,441	77,128	6	9	15
poguda	Low	Target	28,525	7592	25,026	0	2	2
selimo	Low	Target	35,032	12,807	25,129	4	44	48
sojena	Low	Target	20,090	7953	28,256	0	25	25
dekuso	Low	Foil	77,128	14,158	20,090	59	5	64
limasu	Low	Foil	12,807	27,906	8298	46	2	48
mibega	Low	Foil	15,524	19,670	9416	1	45	46
nimola	Low	Foil	12,697	25,129	50,441	155	12	167
posena	Low	Foil	28,525	35,032	28,256	42	29	71
sigume	Low	Foil	30,402	7592	27,117	6	52	58
gosima	Shared	Foil	9502	30,402	27,906	8	347	355
gumeda	Shared	Foil	7592	27,117	25,026	52	181	233
fanibe	Shared	Foil	2995	12,697	19,670	4	383	387
lugaje	Shared	Foil	5496	9416	7953	364	2	366
Modeku	Shared	Foil	25,129	77,128	14,158	238	59	297