Evolutionary account for inter-modality differences in statistical learning

Mikhail Ordin[1,2,*,&], Leona Polyanskaya[1,&], Arthur G. Samuel[1,2,3]

[1]BCBL – Basque Centre on Cognition, Brain and Language, Mikeletegi 69, 20009, San Sebastian, Spain

[2]Ikerbasque – Basque Foundation for Science, Maria Diaz de Hro 3, 48013, Bilbao, Spain

[3]Stony Brook University, Psychology Department, Stony Brook, New York, NY 11794-2500, USA

[&]Equal contribution

[*]corresponding author, address for correspondence: Basque Centre on Cognition, Brain and Language, Paseo Mikeletegi 69, San Sebastian, 20009 Spain. m.ordin@bcbl.eu

Keywords: statistical learning, segmentation of sensory input, evolutionary adaptation of cognitive mechanisms

Evolutionary account for inter-modality differences in statistical learning

Mikhail Ordin, Leona Polyanskaya, Arthur G. Samuel

Significance summary:

Our results show that the efficiency of statistical learning is severely compromised on linguistic material in the visual modality. In the auditory modality, on the contrary, the efficiency of statistical learning is higher on linguistic material. This pattern suggests that the speech faculty in human genus may have been sufficiently long and important for individual fitness to lead to adaptive use of statistical learning mechanisms in the linguistic domain in the auditory modality. Visual language, on the contrary, is a recent cultural invention. There is no evidence of adaptive changes in statistical learning mechanisms operating in the visual modality. This dataset sheds light on cognitive mechanisms underlying the evolution of speech and language faculties, and how these faculties enhanced individual fitness and influenced the same mechanisms which allowed their emergence.

Abstract:

The cognitive mechanisms underlying statistical learning are engaged for the purposes of speech processing and language acquisition. However, these mechanisms are shared by a wide variety of species that do not possess the language faculty. Moreover, statistical learning operates across domains, including non-linguistic material. Ancient mechanisms for segmenting the continuous sensory input into discrete constituents have evolved for general purpose segmentation of the environment, and have been re-adopted for processing linguistic input. Linguistic input provides a rich set of cues to the boundaries between sequential constituents. Such input engages a wider variety of more specialized mechanisms operating on these language-specific cues, thus potentially reducing the role of conditional statistics in tokenizing a continuous linguistic stream. We provide an explicit within-subject comparison of the utility of statistical learning in language vs. non-language domains across the visual and auditory modalities. The results showed that in the auditory modality statistical learning is more efficient on speech-like input, while in the visual modality efficiency is higher on non-language input. We suggest that the speech faculty has been important for individual fitness for an extended period, leading to the adaptation of statistical learning mechanisms for speech processing. This is not the case in the visual modality, in which linguistic material presents a less ecological type of sensory input.

I. Introduction:

The cognitive mechanisms that extract conditional statistics (including transitional probabilities between adjacent elements) are thought to be domain-general(1, 2). It is true that they are used for speech and language processing and acquisition(3), but it appears that they did not evolve for this purpose(4). At least in the auditory modality, they have been shown to operate in species that do not possess the language faculty or even the faculty of vocal learning (5), which is considered to be a pre-requisite for speech (and language) emergence. We suggest that the ancient mechanisms for segmenting the continuous sensory input into discrete constituents have evolved for general purpose segmentation of the environment, and were re-adopted for processing linguistic input.

Speech processing relies on multiple mechanisms and multiple cues, with higher-level cues having more weight for segmenting the continuous input compared to lower-level cues(6). Natural speech, compared to the artificial languages usually used in segmentation experiments, provides more cues and engages more mechanisms in processing of the sensory input. When more cues are available, the relative weighting of each separate cue is diminished(7), with language-specific cues gaining more importance compared to general statistical cues. We

hypothesize that the role of the cognitive mechanisms that extract conditional statistics decreases as the sensory input becomes more language-like. Such input engages a wider variety of more specialized mechanisms operating on language-specific cues, thus reducing the role of conditional statistics in segmentation. Thus, the domain in which we often try to find practical application of these mechanisms is actually not the optimal domain, where these mechanisms function to their full potential.

Tokenizing a continuous sensory input based on conditional statistics has most frequently been studied in the context of speech and language processing and acquisition(8–10). Thiessen et al.(10) provided a comprehensive review of existing theories related to the role of statistical learning mechanisms in splitting a continuous sensory input into discrete constituents, and suggested a broad framework of how statistical learning mechanisms are engaged in segmenting speech into words. They also discussed how different frameworks can be extended beyond word segmentation and be applied, for example, to extracting syntactic constituents in sentence processing.

Despite some evidence favouring cross-modality transfer in statistical learning(11, 12), other studies have systematically failed to find this(13–16). Furthermore, possibly due to individual differences in encoding information in each modality, individual performance both during online testing and offline recognition tests does not correlate across modalities: one individual may perform well in the visual modality, but at chance in the auditory modality, while a different individual may exhibit the reverse pattern(17). Moreover, if two grammars are simultaneously presented in different modalities (e.g., visual and auditory), people can successfully learn both(13), but not when two grammars are presented within the same modality. However, Mitchell and Weiss (12) showed that segmentation of two simultaneously presented streams in different modalities is only possible when the constituent boundaries in the modalities are aligned, thus providing evidence for learning facilitation across modalities.

Although statistical learning appears to be modality-specific, it may be domain-general(1, 18–20). For example, Marcus et al. (21) showed that infants are more successful at extracting rules from sequences of non-linguistic sounds (animal sounds, musical tones, environmental noises) if the same rules had been implemented in speech-like sequences, a transfer effect of acquired statistical regularities across domains.

It has often been suggested that statistical learning mechanisms underlying rule learning are more efficient on linguistic material(20, 21). However, Saffran(9) did not observe a linguistic advantage in extracting and learning rule-based hierarchical relations. In contrast, she did find better learning in the auditory modality than in the visual modality, consistent with what Conway and Christiansen(22) observed. The efficiency of rule extraction and/or learning and generalization can be modulated by the degree of familiarity with a particular type of sensory input(23) and by the ability of this type of sensory information to capture attention(24). Importantly, it is not known whether extraction and generalization of conditional statistical cues operate under the same constraints as rule learning.

The fact that conditional statistics can be efficiently extracted by evolutionarily ancient neural mechanisms(4) that are shared across phylogenetically distant species(5, 25) suggests that statistical learning could have evolved to make sense of the environment(1, 26). If so, such learning might actually be more efficient for non-linguistic sensory input. In the current study, we used an artificial language learning paradigm to test this hypothesis in two different modalities – auditory and visual; auditory language is evolutionarily relative mature, whereas visual language (writing/reading) is quite recent.

In each modality, we developed three stimulus types that differed in how linguistic they were, and tested the operation of statistical learning for each type. In Experiment 1, participants performed a task in the auditory modality. We prepared three types of stimuli: 1) linguistic material, i.e., speech-like sensory input exhibiting linguistic hierarchical structure and acoustic cues manifested in natural languages, in addition to conditional statistics cues; 2) semi-linguistic material, composed of speech-like sub-elements (i.e., syllables), but with acoustic cues that are not typically exploited for linguistic structuring; (3) non-linguistic material, composed of environmental sounds (e.g., water drops, footsteps, animal cries), which had the same conditional statistical cues. The elements in the stream

were organized into recurrent triplets, and we measured how well observers extracted the triplets using a recognition test in which participants were presented with either a triplet or a foil (a tri-syllabic sequence composed of the same syllables or environmental sounds in a sequence that had never occurred consecutively in the familiarization stream).

In Experiment 2, the same participants were exposed to a continuous stream of printed syllables (linguistic material), fractal circles (non-linguistic and non-verbalizable material), and novel complex geometric forms (considered as semi-linguistic because they are not particularly language-like, especially for western participants, yet can be thought of as arbitrary signs of an unknown culture). The stimuli (syllables, circles or figures) were presented sequentially in the center of a screen. The elements in the stream were organized into recurrent triplets, and we measured how well observers extracted the triplets using a recognition test in which participants were presented with either a triplet or a foil. Our two core questions were whether statistical learning operates more/less effectively for linguistic input than for non-linguistic input (a) for evolutionarily mature auditory stimuli, and (b) for more-recent printed stimuli?

II. Method:

II.a. Participants

We recruited 48 native Spanish speakers (18-40 years old) without speech or hearing disorders. They resided in San Sebastian (Spain); most also spoke Basque.

II.b. Material:

EXPERIMENT I: Statistical Learning Efficiency across Domains in the Auditory Modality

We constructed an artificial language learning task(27) to use with three qualitatively different types of stimuli. The properties for these three types are described in detail below. As an overview, one stimulus type was linguistic, one was semi-linguistic, and one was non-linguistic. The **linguistic** stimuli consisted of common Spanish/Basque syllables organized into triplets, with linguistic hierarchical structure and acoustic cues found in natural languages. The **semi-linguistic** stimuli included less common syllables organized into triplets, with acoustic cues that are *not* typically used to produce linguistic structure. The **non-linguistic** stimuli were non-speech environmental sounds, also organized into triplets.

LINGUISTIC INPUT

A continuous stream of syllables was constructed using many presentations of 32 different syllables. One subset of the syllables (24 syllables) was used to create eight "words", each of which was a syllabic triplet: /ko-fa-me/, /fo-na-ku/, /mo-si-ke/, /ka-so-ni/, /sa-mu-pe/, /no-su-pi/, /po-fu-mi/, and /fe-nu-pa/. For these words, the transition probability (TP) from the first syllable (e.g., /ko/) to the second (e.g., /fa/) was 1.0, as was the TP from the second syllable to the third (e.g., /me/). The remaining eight syllables (/ma/, /fi/, /pu/, /se/, /ne/, /ki/, /li/, and /lu/) were used to make the syllabic stream more language-like, by separating the three-syllable words with single-syllable filler sounds (like most function words in a language). This was a modification of an approach by Gervain et al.(28), who introduced frequent syllables to model function words, thus making the input more language-like.

The structure of the syllabic stream is represented in Figure 1. Orange squares represent syllabic triplets and blue squares represent filler syllables. Triplets model content words and fillers model high-frequency non-content words such as prepositions, articles, and morphological elements. While the TPs between adjacent syllables within triplets were 1.0, the TPs between filler syllables and syllables straddling the triplet boundaries were approximately 0.125.

The difference in TPs between syllables within triplets (whole constituents embedded in the continuous input) and TPs for syllables straddling the boundaries of the holistic units, provides a potential statistical basis to extract the triplets from the input. Each triplet was presented 80 times, with the syllable concatenation constrained to not produce real words of Spanish or Basque.

INSERT FIGURE 1 SOMEWHERE HERE

The triplets with their surrounding filler syllables represent phonological phrases (PPs). As in natural languages, they were defined by intonational contours with rising initial boundary tones and either rising or falling final boundary tones. Pairs of PPs were united into intonational phrases (IPs). This patterning imitates the hierarchical prosodic structuring that is typical in natural languages(29), using acoustic cues that are used for prosodic structuring in natural languages(30). PPs and IPs were defined by boundary tones imposed on the phrase-initial and phrase-final syllables, and by pauses (longer between IPs than between PPs within IPs). Triplet edges were not aligned with the edges of PPs (PP-initial and PP-final elements were represented by fillers), which did not allow using boundary cues for extracting them from a continuous input. Figure 2 shows a waveform, spectrogram and intonational contour of a PP pair within one IP.

INSERT FIGURE 2 SOMEWHERE HERE

MBROLA was used to synthesize the speech. We used the SP2 voice (Spanish), which produces phonemes in a Spanish-specific manner, making the input more speech-like for our participants.

SEMI-LINGUISTIC INPUT

For this type of input, we used a different set of syllables, including phonemes that are rare or non-existent in Spanish and Basque. Again, 24 syllables were used to create 8 triplets (/ba-go-dzi/, /gu-be-llu/, /to-tzu-di/, /ta-tse-ga/, /dzo-da-tu/, /de-bi-lla/, /lle-du-bo/, /dza-gi-llo/) and 8 additional syllables were used for the fillers (/do/, /ge/, /bu/, /ti/, /te/, /lli/, /sha/, /sho/) (/ll/ is a palatal lateral sonorant, /ts/ and /dz/ are voiceless and voiced dental alveolar affricates). Syllable and pause durations were equal to those used in linguistic stimuli. The stimuli were synthesized using the Italian IT3 voice in MBROLA in order to make the phoneme realizations less familiar to our participants. Also, we inserted 1-ms pauses between each syllable. These pauses are not detectable by the human auditory system, but they blocked co-articulatory transitions between syllables, which were synthesized by the algorithm as if each syllable was pronounced isolation. The syllabic stream was synthesized with a constant (and thus unnatural) pitch (120Hz), and boundary cues were implemented by linear intensity ramping instead of the pitch and timing cues that are typical in natural language. We applied amplitude ramping on the edges of IPs (over 2 syllables) and PPs within IPs (over 1 syllable). Each triplet was presented 80 times.

The stimuli were designed to be semi-linguistic because they included (linguistic) syllables, but also included features that were unnatural for the listeners. These included the use of intensity fluctuations to mark prosodic boundaries, non-native realizations of atypical phonemes, and a lack of co-articulatory transitions. As described below, a test was implemented to verify that these stimuli were indeed perceived to be less linguistic than the “Linguistic” stimuli.

NON-LINGUISTIC INPUT

To create non-linguistic auditory input, we selected 32 natural environmental sounds (water drops, footsteps, squeaks, animal noises, etc.) from https://freesound.org. All sounds were equalized in duration to 300ms. Pause durations were equal to those used in linguistic stimuli. All sounds were normalized in intensity to 80dB. 24 sounds were used for triplets and 8 sounds for fillers, using the same statistical structure implemented in linguistic and semi-linguistic stimuli. All the sounds had their amplitude normalized to the same average – to prevent some sounds

being louder than others – and then concatenated into a continuous stream in PRAAT(31). Ramping (using the same approach implemented in semi-linguistic stimuli) was used for hierarchical structuring (see Figure 3). Each triplet was presented 80 times. Short samples of linguistic, semi-linguistic and non-linguistic material can be found in supplementary material online.

INSERT FIGURE 3 SOMEWHERE HERE

EXPERIMENT II: Statistical Learning Efficiency across Domains in the Visual Modality

As in the auditory modality, linguistic, semi-linguistic and non-linguistic continuous sensory inputs were prepared in the visual modality. Each contained recurrent triplets embedded into streams with filler elements. In the visual modality the construction of the stimuli was similar to that in the auditory modality, but several parameters were adjusted to accommodate differences in processing abilities in the two modalities.

In the auditory modality statistical learning is more attuned to temporal regularities, and in the visual modality – to spatial regularities(32). As we only had temporal regularities in our design, we had to make complexity in the visual modality lower, in order to accommodate this difference. Also, the rate of presentation in the visual modality was slower because at the rate at which the sounds were presented, the TPs between images are not registered. Overall, statistical learning in the visual modality is efficient only at a slow rate, whereas efficiency is not strongly modulated by the rate of presentation of sequential elements in the auditory modality(33). As the presentation rate in the visual modality was slower, we had to reduce the number of triplets (which in turn reduces complexity); this also made the duration of the familiarization stream reasonable (otherwise, the visual stream would be too long). At the same time, we kept the TPs and the overall structure of the input equivalent across modalities (triplets were embedded into larger constituents, with fillers at the edges of the larger constituents). These methodological decisions were made to equilibrate statistical learning tasks across modalities, yet fundamental differences between modalities make it difficult to compare efficiency of statistical learning across modalities directly. Therefore, the analysis will focus of direct comparison between domains (linguistic vs. non-linguistic) within modalities, and on subject-based correlations of individual performance in different modalities and domains (in order to explore whether higher efficiency in one modality/domain is associated with a higher efficiency on a different modality/domain), without a direct across-modality comparison at the group level.

LINGUISTIC INPUT

For linguistic input, we selected 20 syllables. Of these, 12 syllables were used for recurrent triplets (TE-GU-BA, TA-BO-FA, KA-BE-TO, GA-FO-BU), and 8 syllables were used for fillers (TU, GO, GE, KO, KU, FU, FE, KE). All stimuli were presented in upper case letters. Note that these syllables were different from those used in the auditory modality. The statistical structure of the visual input matched that implemented in the auditory modality (TPs set to 1.0 within triplets and around .0125 at the boundary of the constituents). In the stream, each triplet was presented 50 times. Instead of pauses, we used commas (within IPs) and dots (between IPs). The syllables were presented in the middle of the screen, one by one, and each syllable or punctuation mark stayed on the screen for 500ms. This duration is longer than the duration of syllables or sounds in the auditory modality.

SEMI-LINGUISTIC INPUT

For the semi-linguistic input, we used 20 geometric shapes organized into four triplets and eight fillers (see Figure 4b) with the same statistical structure as syllables in the linguistic condition. Again, each triplet was presented 50 times. For the boundary signals, we used grey squares (instead of commas) and white squares (instead of dots), each also presented for 500ms. Subjectively, white squares looked more prominent than grey squares, which was one tone lighter than the grey background, against which the shapes were displayed.

NON-LINGUISTIC INPUT

For non-linguistic stimuli, we used fractals generated at http://sirxemic.github.io/ifs-animator/, and created continuous visual input with a statistical structure identical to that implemented for the semi-linguistic and linguistic visual input (see Figure 4a).

We assumed that participants could potentially accept geometrical forms as symbols able to convey a message in a linguistic way (the assumption was later verified in a post-experiment test), as these are somewhat like hieroglyphs or cuneiforms. While these not familiar to our participants, they are elements of writing systems for natural languages. Elements like fractals have not been observed in human languages as linguistic symbols. As such, they are at the end of the continuum on a linguistic - non-linguistic spectrum.

INSERT FIGURES 4a and 4b SOMEWHERE HERE

II. c. PROCEDURE:

People came to the laboratory twice, with an interval of at least one week between the sessions. Each time they performed the experiment in one of the modalities. During each session, participants were tested with linguistic, semi-linguistic and non-linguistic stimuli (in a counter-balanced order). The order of modalities was counterbalanced across participants.

In the auditory modality, participants were told that they would have to train their ear to detect the recurrent sequences of sounds or syllables in the auditory stream, and that afterwards they would have to recognize the embedded sequences during a test. In the visual modality, participants were told that they would have to train their eyes to detect the recurrent sequences of images or syllables, presented one by one in the center of the screen, and that afterwards they would have to recognize the embedded sequences during a test. It is important to note that statistical learning is both intentional and automatic (34). Explicit instructions turn automatic processing into intentional learning, which requires attention to what is to be learnt. In statistical learning, attention itself is modulated as learning progresses(35, 36). Attention influences learning by facilitating encoding of particular aspects of the input (and explicit instructions provide participants with information about what aspects of the input they need to pay attention to – recurrence of sequences within a continuous stream). At the same time, learning affects attention by creating perceptual salience of such regularities, and drawing attention to violations of these regularities. Importantly, the question of whether explicit versus implicit instructions affect performance during the recognition test is still debated. Although some researchers found statistical learning to be more efficient when instructions are explicit(37–40), others did not observe performance differences between explicit and implicit conditions(41, 42). In our study, the choice of explicit conditions was determined by the within-subject design of the experiment. When participants receive implicit instructions for one type of familiarization stream, they will be consciously aware of what to look for after the test, and the type of instructions will not matter. Thus we decided to avoid this by giving participants explicit instructions from the start.

Presentation of each type of continuous input was followed by a recognition test. In the auditory modality, on the recognition test participants listened to a triplet of syllables (F0 set to 120Hz throughout) or environmental sounds (normalized for intensity), with syllable/sound durations corresponding to those heard during the learning phase. For the test items, we presented either actual triplets, or foils. The syllables or sounds in the foils were ones that were used in actual triplets, but no actual triplets included two elements of any foil (i.e., both adjacent and non-adjacent TPs were zero). The foils were of two types. In type-1 foils (*ordered foils*), we preserved the order of elements (syllables or sounds). For example, if a particular syllable was used in the triplet-initial position in a learning stream, it was also used in the foil-initial position. Type-2 foils (random foils) violated the ordinal position of elements inside the triplets, and thus a triplet-initial element (during training) was presented in foil-medial or foil-final position in

type-2 foils only, avoiding positional matches. We created 8 random and 8 ordered foils. Participants were asked to listen to a test item and to indicate whether the item had been embedded in the continuous stream of syllables/sounds as a whole sequence (binary choice), and to indicate whether they were sure in their response (confidence ratings were collected for a different project and were not analyzed here).

For the auditory modality, triplets and foils were used twice during the test, yielding 48 trials per stimulus type (linguistic, semi-linguistic and non-linguistic). The recognition test followed the learning stream immediately. After the test, participants were offered a short break before the next continuous input was presented for learning. After the third recognition test, participants listened to six 15-second passages from the learning streams, 2 passages per each type. They judged which sounds seemed most speech-like and which seemed least speech-like.

Similarly, we created 4 random and 4 ordered foils in the visual modality per stimulus type. Each type of learning input was followed by a test with 24 trials. During the test, participants looked at either actual triplets extracted from the learning stream or foils; for each, they indicated whether the item had been embedded in the learning stream or not.

After the third recognition test, a norming test was run to confirm that linguistic material was indeed perceived by participants as most speech- or language-like, non-linguistic material was perceived as least speech- and language-like, and semi-linguistic material was perceived as intermediate between the two extremes. Participants watched six streams of alternating images from the learning phase, 2 per stimulus type, for 15 seconds each. They judged which type looked most language-like and which least language-like. In both modalities, all participants indicated that our linguistic stimuli looked or sounded the most language-/speech-like, and non-linguistic stimuli the least language-/speech-like.

III. Analysis:

We calculated the sensitivity index (D') by considering the triplets correctly endorsed as sequences from the familiarization stream as *hits* and the foils (irrespective of whether the ordinal position of elements in the foils was maintained or not) as *false alarms*. Consequently, rejected foils were considered as *correct rejections*, and rejected triplets as *misses*. In the Appendix, we provide the analysis of D' for triplets relative to ordered foils and random foils separately. However, foil type made no difference in the result pattern, and therefore we do not report these analyses in the main manuscript.

D' and bias (C) measures were calculated separately for linguistic, semi-linguistic and non-linguistic stimuli in each modality (Figure 5a for D' and 5b for C). Given the necessary differences across modalities in the amount of exposure, task difficulty (number of triplets) and different underlying cognitive mechanisms, the analyses focused on comparing sensitivity and bias across domains within each modality.

INSERT FIGURES 5a and 5b SOMEWHERE HERE

Sensitivity Analyses: The analyses were performed by constructing linear mixed models (in SPSS v.18) with *subject* as a random factor and *domain* (linguistic vs. semi-linguistic vs. non-linguistic) as a fixed factor. Models with different covariance types were constructed (including AR(1), Diagonal, Compound Symmetry, Scaled Identity and Unstructured types). The reported model (with Scaled Identity) was chosen as the one with the minimum parameters (n=5) and lowest 2LL information criterion. The model with Unstructured covariance type failed to converge and was thus not considered. Parameter estimation was done by a restricted maximum likelihood algorithm (threshold of 1000 iteration for model fitting).

In the auditory modality, the effect of domain was significant, F(2, 94)=5.469, p=.006. For linguistic material, β=.29(SE=.09), t(94)=3.27, p=.002 was larger than for non-linguistic material, β=.11(.09), t(94)=1.2, p=.233. D' was higher for linguistic and non-linguistic material than for semi-linguistic material, with a larger difference between linguistic and semi-linguistic than between non-linguistic and semi-linguistic stimuli. Paired t-tests (two-tailed, reported after Bonferroni correction) confirmed that sensitivity for semi-linguistic material (D'=.23, SD=.63) was significantly lower than on linguistic material, t(47)=3.337, p=.004, M=-.29, 95%CI[.116:.468]. Sensitivity was higher in the linguistic (D'=.52, SD=.66) than in the non-linguistic domain (D'=.34, SD=.61), although the difference was not significant, t(47)=2.164, p=.072, M=-.18, 95%CI[.013:.357].

In the visual modality, the effect of domain was significant, F(2, 94)=5.469, p=.006. For linguistic material, β=-.25(.15), t(94)=-1.65, p=.102 was lower than for non-linguistic material, β=.16(.15), t(94)=1.076, p=.285. D' was the highest on non-linguistic and lowest on linguistic material. Paired t-tests (2-tailed, corrected p values are reported) confirmed that in the visual modality, sensitivity in the linguistic domain (D'=.25, SD=.59) was significantly lower than in the non-linguistic domain (D'=.67, SD=.89), t(47)=-2.609, p=.024, M=-.42, 95%CI[-.744:-.096]. The difference in sensitivity between linguistic and semi-linguistic (D'=.5, SD=.87) material in the visual modality was not significant, t(47)=-1.541, p=.26, M=-.25, 95%CI[-.586:.0777].

Overall, the results demonstrate that in the visual modality, statistical learning is more efficient for non-linguistic than for linguistic sensory input. In the auditory modality a reverse pattern is observed: efficiency is higher for linguistic than non-linguistic sensory input. The very poor performance on semi-linguistic material in the auditory modality might seem puzzling. However, it can be explained by the nature of these stimuli. Being composed of syllables, semi-linguistic stimuli may recruit mechanisms that are brought to bear while listening to natural speech, but these mechanisms are useless for material intentionally designed to be unprocessable as linguistic input (recall that these stimuli are composed of unusual sounds and inappropriate prosodic cues, which mismatch listeners' expectations, and could even be argued to be closer to non-linguistic than linguistic material). Recruitment of speech processing mechanisms thus disrupts their processing rather than facilitates it).

We calculated Pearson correlations (all p-values are two-tailed) for D' scores over subjects across modalities and across domains. We did not observe correlations of D' scores across modalities for linguistic (r=.06, p=.678), semi-linguistic (r=-.074, p=.619) or non-linguistic (r=.111, p=.452) material. This is consistent with an extensive prior literature (see (17) for an overview) that suggests a cross-modality barrier in the transfer of statistical learning. Interestingly, we found a strong correlation of D' scores on linguistic vs. non-linguistic material (r=.572, p<.0005) in the auditory modality, but not in the visual modality (r=-.096, p=.516). The difference in correlation strength of D' scores across domains within modalities (r=-.096 and r=.572) is itself significant (z=3.542, p<.0005). Although statistical learning is often claimed to be modality-specific yet domain-general (Frost et al., 2015), we observed clear evidence for the former, and a more nuanced interpretation for the latter (domain-generality in the auditory but not in visual modality). The reliable correlation in the auditory modality likely reflects memory span differences among individuals. As auditory statistical learning is more tuned for processing sequential input, it may rely on echoic memory to keep auditory chunks ready for further processing. As noted above, visual statistical learning is more attuned to spatial regularities(13).

Control for Bias:

For the bias measure (C), the effect of *domain* was significant in the auditory modality, F(2,94)=12.172, p<.0005. Estimated β=.077(.06), t(94)=1.234, p=.22 for linguistic material and β=.3(.06), t(94)=4.754, p<.0005 for non-linguistic material showed a stronger tendency to accept non-linguistic tokens than on semi-linguistic and linguistic tokens. In the visual modality, the effect of *domain* was not significant, F(2,94)=.462, p=.631. Estimated β=.0036(.078), t(94)=.045, p=.964 for linguistic material and β=.067(.078), t(94)=.855, p=.395 showed that the bias to accept or reject presented tokens was not different between stimuli types. Importantly, we calculated Pearson correlations over subjects for bias measures across domains within modalities and across modalities within domains. All of the

correlations were significant (all p-values are below .001) and substantial (.388<r<.585), suggesting that if a participant tends to reject or accept test items in one modality/domain, that participant applies the same strategy for the other modalities / stimulus types.

IV. Discussion:

The current study provides an explicit, within-subject comparison of the utility of statistical learning in linguistic vs. non-linguistic domains in the visual and auditory modalities. Overall, the data showed that in the visual modality, statistical learning is more efficient for non-linguistic material, whereas in the auditory modality statistical learning was more effective for linguistic sounds. Individual performance was not correlated across modalities, consistent with prior research. Within modalities but across domains (language vs. non-language), performance was not correlated in the visual modality, while being strongly correlated in the auditory modality. This difference in correlations within modalities across domains was large and significant.

This finding can be accounted for by considering historical differences between spoken and written language faculties. Despite a heated debate about when the speech faculty first emerged in the human genus, with estimates as short as 100,000 years(43) and as long as 2,000,000 years(44, 45), with a range of potential timepoints in-between(46), it is clear that spoken language considerably preceded the invention of writing. Written language is a much more recent cultural invention, dating back to the early Neolithic period, around the 7th millennium B.C.(47, 48). This early form of proto-writing was a hieroglyphic script, and even contemporary scripts of this type often represent whole concepts as orthographic entities (47, 48). The predictive power of different concepts is akin to the predictive power of one word to predict the following word, or the constituent-final syllable to predict the initial syllable of the following constituent. In the auditory modality, word boundaries are signaled by a dip in the stream of transitional probabilities(49). Consequently, the role of conditional statistics in processing visual language gained in importance only when syllabic/alphabetic writing was invented around the 5th century B.C. Moreover, the spread of reading and writing skills within populations was quite slow, further limiting the window of opportunities for efficiency of processing visual language to impact individual fitness. Consequently, the efficiency of statistical learning for language in the visual modality has not had very long to develop, as it has in the auditory modality. Speech provided a much wider window of opportunity for the adaptation of statistical learning mechanisms. Over time, speech, being pervasive and important, gained precedence and re-tuned these mechanisms from processing environmental sounds to speech.

Statistical learning in the visual modality on linguistic material may be affected by reading and writing acquisition, and individual variation in reading skills. Statistical learning of non-linguistic material is shaped more by phylogenetic rather than ontogenetic influences. In the auditory modality, statistical learning both of speech-like material and of environmental sounds is influenced by phylogenetic factors. This is reinforced by the correlational analyses and provides additional support for the hypothesis that statistical learning in the auditory modality has been shaped by the speech faculty in the human genus and re-adapted for speech processing. In the visual modality, in contrast, statistical learning is not attuned for processing written text. Such text has not been an ecologically natural sensory input in the environment of the *homo sapiens sapiens* species for the very long time scale needed for evolutionary adaptation, and thus core statistical learning mechanisms have not (yet) been adapted for written input.

An alternative hypothesis, which is also supported by the data and is based on historical differences between speech and written language faculties, is grounded in the effect of cultural evolution of the design of the communicative systems. It has been proposed that neither speech(50) nor writing/reading(51) benefits from evolutionary specialization for processing linguistic material. Enhanced efficiency of statistical learning mechanisms on speech-like stimuli is explained by the pressure of cultural selection on maintaining properties that are more easily processed by available neural circuitry and cognitive resources. Those aspects of signals that afford easier processing by available resources facilitate communication and are more easily transferred across generations and thus become more stable

in populations (50, 52, 53). As vocal communication has been subject to cultural selection for a longer time, the existing patterns of the vocal communicative system are more prone to be processed by existing mechanisms than environmental sounds. This argument suggests that it is not speech that affects statistical learning mechanisms in the auditory modality via *natural* selection, but rather those speech patterns which are processed more efficiently are then passed through generations by means of *cultural* evolution. Cultural evolution is a fast-paced process (sometimes requiring only several generations), and the effect of cultural selection should also be observed on linguistic material in the visual modality. The effect of enculturation via written language can be observed even at the timescale of individuals: literacy influences cortical maps and leads to re-organization of cognitive processing in ontogenesis(51, 54). Thus, we would need to understand why statistical learning in the visual modality is less prone to cultural selection pressure than in the auditory modality, or admit that the data is better accounted for by selection pressure on the cognitive system.

Importantly, we do not postulate the existence of speech-specific cognitive or neural modules that have emerged under the pressure of natural selection. Rather, we are talking about an elevated recycling of existing neural circuitry and cognitive machinery, when speech has become essential for individual fitness. Re-adaptations of the existing machinery to process (and produce) speech-like input over the long term (at the time-scale of humanity) is supported by the fact that rapid presentation of non-speech sounds, at the rate of syllables, leads to merging these sounds into an acoustic stream of unrecognizable sounds and/or inability to perceive their sequential order. However, presentation of syllables at the same rate is perceived as intelligible speech (50, 55). At the same time, cognitive processes operating over long stochastic (statistically predictable but not deterministic) sequences of non-linguistic sounds and sequences of speech units (with corresponding durations) are similar (56), as is the underlying neural machinery(50). This suggests that long-term experience with a particular type of input leads to enhanced processing of this type.

In the design of the current study, the auditory modality potentially had an advantage because we explored statistical learning in sequential presentation (i.e., in time); visual statistical learning works better in space(32). It is difficult to predict the results that might be obtained if both visual and auditory stimuli presentation were manipulated spatially. When several simultaneous auditory inputs are present (e.g., think of a party conversation, surrounded by many other interacting people), humans focus on a particular source, and focus on how it develops in time. In the visual modality, on the contrary, spatial patterns tend to be extracted.

We should also note that the language relevance of our material is correlated with familiarity, and thus with ease of encoding. Linguistic stimuli are more familiar because people are more often exposed to streams of syllables than to streams of unrelated environmental sounds. Within the evolutionary perspective that we take, the development of any ability can be thought of as a response to a frequent need, so the evolutionary view is framed as being frequency-sensitive over a very long time (if a stimulus is rarely encountered, why would an ability to process it develop?) It is important is to separate very long term frequency effects (evolution) from just long term ones (i.e., within the individual's lifetime). On the timescale of humanity, the amount of exposure to spoken language is much larger than to written language, and this can explain adaptations in the auditory modality but not in the visual modality. From an ontogenetic perspective, in both modalities, individuals accumulate larger exposure to and thus familiarity with linguistic input. On these grounds, one would expect enhancement in statistical learning in both modalities due to ontogenetic factors, versus only in the auditory modality due to evolutionary adaptations. Our data thus highlight the evolutionary adaptation of statistical learning mechanisms on the timescale of humanity.

References:

1. N. Z. Kirkham, J. A. Slemmer, S. P. Johnson, Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition* **83** (2002).

2. E. D. Thiessen, Domain General Constraints on Statistical Learning. *Child Dev.* **82**, 462–470 (2011).

3. L. C. Erickson, E. D. Thiessen, Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Dev. Rev.* **37**, 66–108 (2015).

4. Y. Kikuchi, W. Sedley, T. D. Griffiths, C. I. Petkov, Evolutionarily conserved neural signatures involved in sequencing predictions and their relevance for language. *Curr. Opin. Behav. Sci.* **21**, 145–153 (2018).

5. A. E. Milne, C. I. Petkov, B. Wilson, Auditory and Visual Sequence Learning in Humans and Monkeys using an Artificial Grammar Learning Paradigm. *Neuroscience* **389**, 104–117 (2018).

6. S. L. Mattys, L. White, J. F. Melhorn, Integration of multiple speech segmentation cues: A hierarchical framework. *J. Exp. Psychol. Gen.* **134**, 477–500 (2005).

7. S. L. Mattys, J. Brooks, M. Cooke, Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cogn. Psychol.* **59**, 203–243 (2009).

8. Y.-H. Hung, S. J. Frost, K. R. Pugh, "Domain Generality and Specificity of Statistical Learning and its Relation with Reading Ability" in (2018), pp. 33–55.

9. J. R. Saffran, Constraints on statistical language learning. *J. Mem. Lang.* **47**, 172–196 (2002).

10. E. D. Thiessen, A. T. Kronstein, D. G. Hufnagle, The extraction and integration framework: A two-process account of statistical learning. *Psychol. Bull.* **139**, 792–814 (2013).

11. G. T. M. Altmann, Z. Dienes, A. Goode, "Modality Independence of Implicitly Learned Grammatical Knowledge" (1995).

12. A. D. Mitchel, D. J. Weiss, Learning Across Senses: Cross-Modal Effects in Multisensory Statistical Learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 1081–1091 (2011).

13. C. M. Conway, M. H. Christiansen, Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychol. Sci.* **17**, 905–912 (2006).

14. R. L. Gomez, L. Gerken, R. W. Schvaneveldt, The basis of transfer in artificial grammar learning. *Mem. Cogn.* **28**, 253–263 (2000).

15. M. Redington, N. Chater, Transfer in Artificial Grammar Learning: A Reevaluation. *J. Exp. Psychol. Gen.* **125**, 123–138 (1996).

16. R. J. Tunney, G. T. M. Altmann, The Transfer Effect in Artificial Grammar Learning: Reappraising the Evidence on the Transfer of Sequential Dependencies. *J. Exp. Psychol. Learn. Mem. Cogn.* **25**, 1322–1333 (1999).

17. N. Siegelman, L. Bogaerts, M. H. Christiansen, R. Frost, Towards a theory of individual differences in statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* **372** (2017).

18. R. Frost, B. C. Armstrong, N. Siegelman, M. H. Christiansen, Domain generality versus modality specificity: The paradox of statistical learning. *Trends Cogn. Sci.* **19**, 117–125 (2015).

19. A. L. Gebhart, E. L. Newport, R. N. Aslin, Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychon. Bull. Rev.* **16**, 486–490 (2009).

20. E. D. Thiessen, Effects of Inter- and Intra-modal Redundancy on Infants' Rule Learning. *Lang. Learn. Dev.* **8**, 197–214 (2012).

21. G. F. Marcus, K. J. Fernandes, S. P. Johnson, Infant Rule Learning Facilitated by Speech. *Psychol. Sci.* **18**, 387–391 (2007).

22. C. M. Conway, M. H. Christiansen, Modality-constrained statistical learning of tactile, visual, and auditory sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 24–39 (2005).

23. C. Dawson, L. A. Gerken, From domain-generality to domain-sensitivity: 4-Month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition* **111**, 378–382 (2009).

24. A. Vouloumanos, J. F. Werker, Tuned to the signal: The privileged status of speech for young infants. *Dev. Sci.* **7**, 270–276 (2004).

25. T. Q. Gentner, K. M. Fenn, D. Margoliash, H. C. Nusbaum, Recursive syntactic pattern learning by songbirds. *Nature* **440**, 1204–1207 (2006).

26. M. Ordin, L. Polyanskaya, D. Soto, Neural bases of learning and recognition of statistical regularities. *Ann. N. Y. Acad. Sci.*, nyas.14299 (2020).

27. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants. *Science (80-.).* **274**, 1926–1928 (1996).

28. J. Gervain, M. Nespor, R. Mazuka, R. Horie, J. Mehler, Bootstrapping word order in prelexical infants: A Japanese-Italian cross-linguistic study. *Cogn. Psychol.* **57**, 56–74 (2008).

29. M. Nespor, I. Vogel, *Prosodic Phonology* (DE GRUYTER, 2007) https:/doi.org/10.1515/9783110977790.

30. C. Gussenhoven, *The phonology of tone and intonation* (Cambridge University Press, 2004).

31. P. Boersma, Praat, a system for doing phonetics by computer. *Glot Int.* **5** (2002).

32. C. M. Conway, How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neurosci. Biobehav. Rev.* **112**, 279–299 (2020).

33. C. M. Conway, M. H. Christiansen, Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *Eur. J. Cogn. Psychol.* **21**, 561–580 (2009).

34. N. B. Turk-Browne, J. A. Jungé, B. J. Scholl, The automaticity of visual statistical learning. *J. Exp. Psychol. Gen.* **134**, 552–564 (2005).

35. A. Alamia, A. Zénon, Statistical Regularities Attract Attention when Task-Relevant. *Front. Hum. Neurosci.* **10**, 42 (2016).

36. B. M. Hard, M. Meyer, D. Baldwin, Attention reorganizes as structure is detected in dynamic action. *Mem. Cognit.* **47**, 17–32 (2019).

37. S. Kahta, R. Schiff, Implicit learning deficits among adults with developmental dyslexia. *Ann. Dyslexia* **66**, 235–250 (2016).

38. M. Laasonen, *et al.*, Project DyAdd: Implicit learning in adult dyslexia and ADHD. *Ann. Dyslexia* **64**, 1–33 (2014).

39. A. S. Reber, F. F. Walkenfeld, R. Hernstadt, Implicit and Explicit Learning: Individual Differences and IQ. *J. Exp. Psychol. Learn. Mem. Cogn.* **17**, 888–896 (1991).

40. R. Schiff, A. Sasson, G. Star, S. Kahta, The role of feedback in implicit and explicit artificial grammar learning: a comparison between dyslexic and non-dyslexic adults. *Ann. Dyslexia* **67**, 333–355 (2017).

41. J. Arciuli, J. von K. Torkildsen, D. J. Stevens, I. C. Simpson, Statistical learning under incidental versus intentional conditions. *Front. Psychol.* **5**, 747 (2014).

42. Z. Dienes, D. Broadbent, D. Berry, Implicit and Explicit Knowledge Bases in Artificial Grammar Learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **17**, 875–887 (1991).

43. P. Mellars, Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9381–9386 (2006).

44. A. Belfer-Cohen, N. Goren-Inbar, Cognition and Communication in the Levantine Lower Palaeolithic. *World Archaeol.* **26**, 144–157.

45. F. d'Errico, *et al.*, Archaeological evidence for the emergence of language, symbolism, and music - An alternative multidisciplinary perspective. *J. World Prehistory* **17**, 1–70 (2003).

46. T. J. H. Morgan, *et al.*, Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nat. Commun.* **6**, 1–8 (2015).

47. H. Stephen D., *The First Writing: Script Invention as History and Process.* (Cambridge University Press, 2004).

48. M. Gross, The evolution of writing. *Curr. Biol.* **22**, R981–R984 (2012).

49. M. Ordin, L. Polyanskaya, D. Soto, N. Molinaro, Electrophysiology of statistical learning: Exploring the online learning process and offline learning product. *Eur. J. Neurosci.*, ejn.14657 (2020).

50. W. T. Fitch, The Biology and Evolution of Speech: A Comparative Analysis. *Annu. Rev. Linguist.* **4**, 255–279 (2018).

51. S. Dehaene, L. Cohen, Cultural recycling of cortical maps. *Neuron* **56**, 384–398 (2007).

52. T. L. Griffiths, M. L. Kalish, S. Lewandowsky, Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3503–3514 (2008).

53. K. Smith, S. Kirby, Cultural evolution: implications for understanding the human language faculty and its evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3591–3603 (2008).

54. S. Dehaene, *et al.*, How learning to read changes the cortical networks for vision and language. *Science (80-.).* **330**, 1359–1364 (2010).

55. R. M. Warren, C. J. Obusek, R. M. Farmer, R. P. Warren, Auditory Sequence: Confusion of Patterns Other Than Speech or Music. *Science (80-.).* **164**, 586–587 (1969).

56. R. M. Warren, J. A. Bashford, J. M. Cooley, B. S. Brubaker, Detection of acoustic repetition for very long stochastic patterns. *Percept. Psychophys.* **63**, 175–182 (2001).

/ma pu **ka—so—ni** se **sa—mu—pe** lu fi **no—su—pi** ne li **po—fu—mi** ki li/

Figure 1. A schematic representation of a syllabic stream, showing 4 triplets (orange squares, bold font), with fillers (blue squares, normal font), and a 50-ms within-IP between PPs pause.

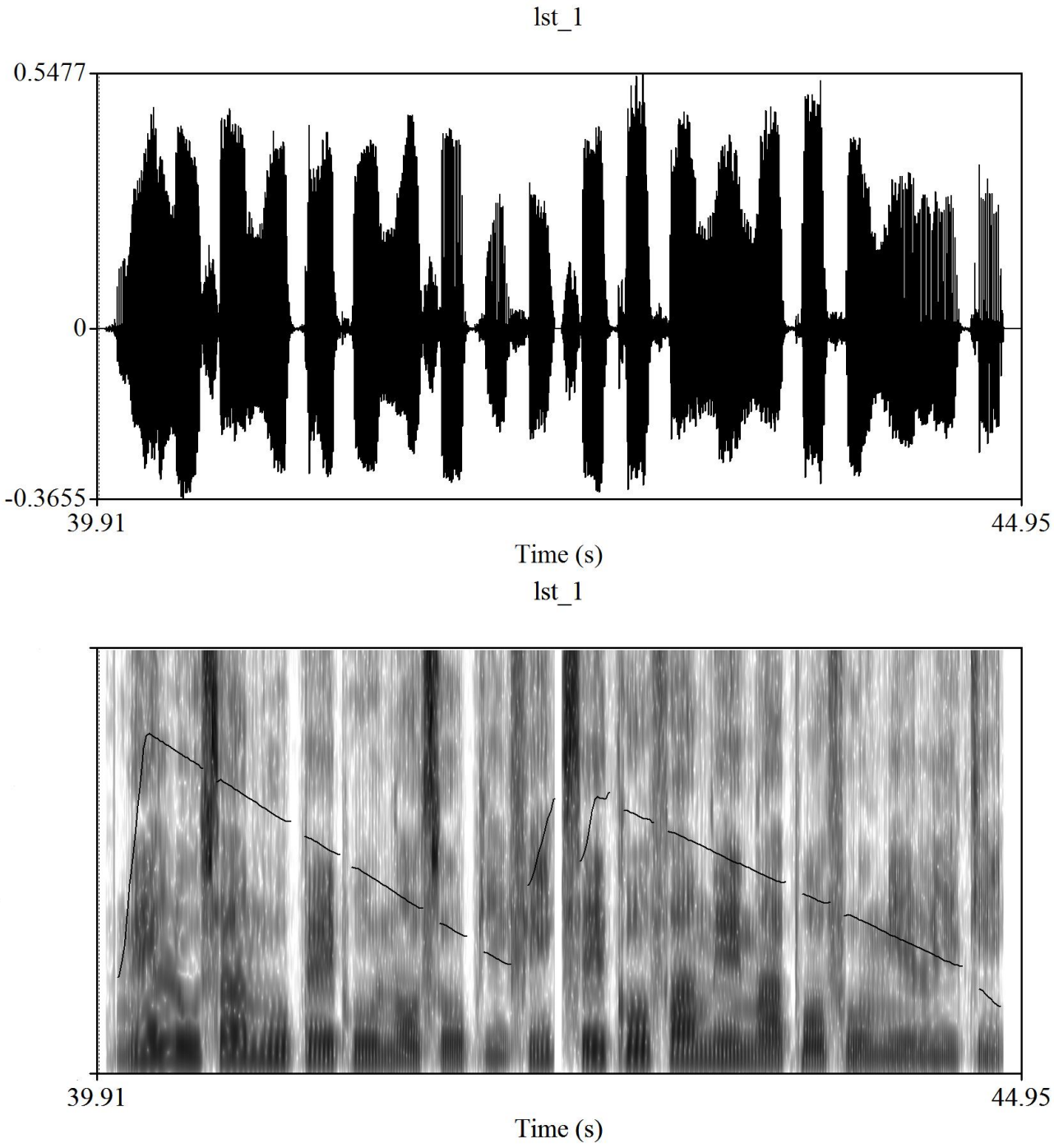


Figure 2. Waveform and spectrogram representing one IP, with a 50-ms pause between two PPs, defined by boundary tones. The spectrogram is displayed on the scale between 0 and 7000Hz, and the pitch contour is displayed on a scale between 50 and 250Hz.

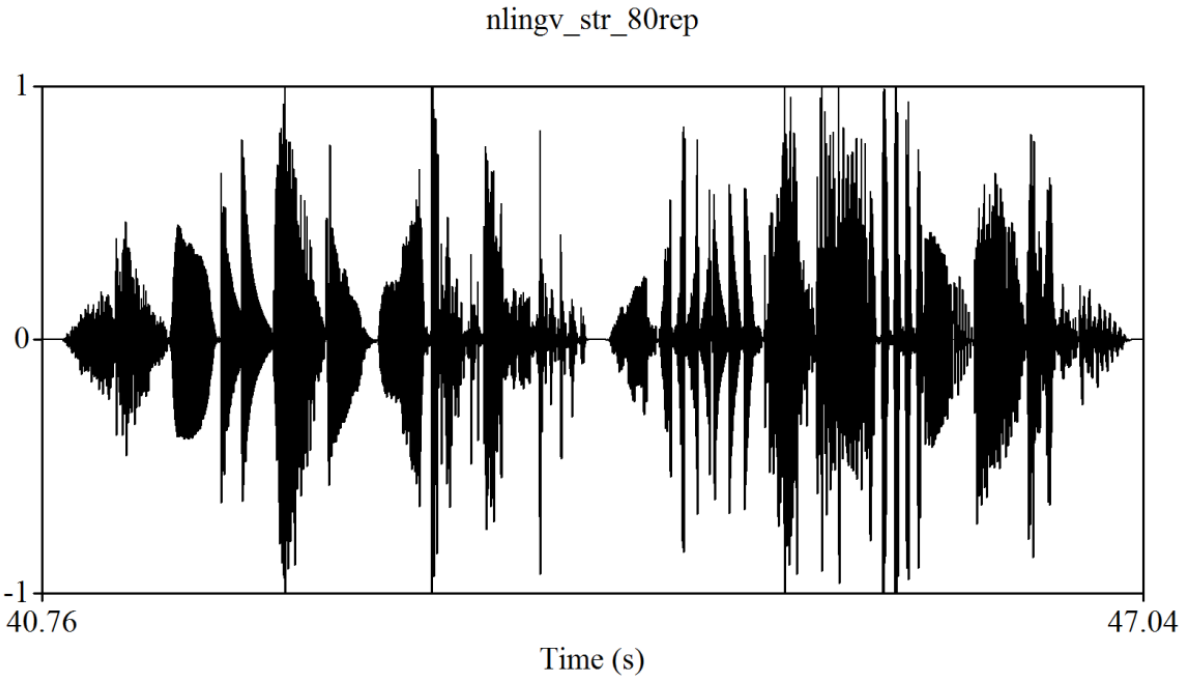


Figure 3. Waveforms of non-linguistic stimuli, showing intensity ramping between PPs and IPs used to define the hierarchical structure.

EXPERIMENT II: Statistical Learning Efficiency across Domains in the Visual Modality

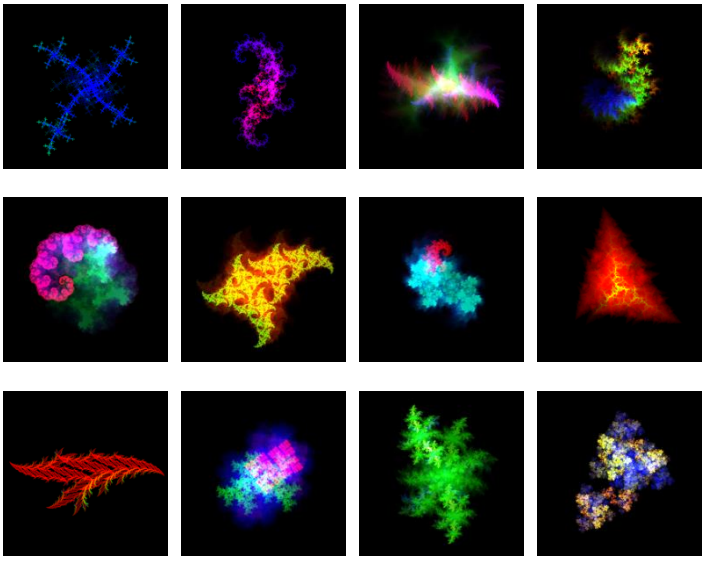
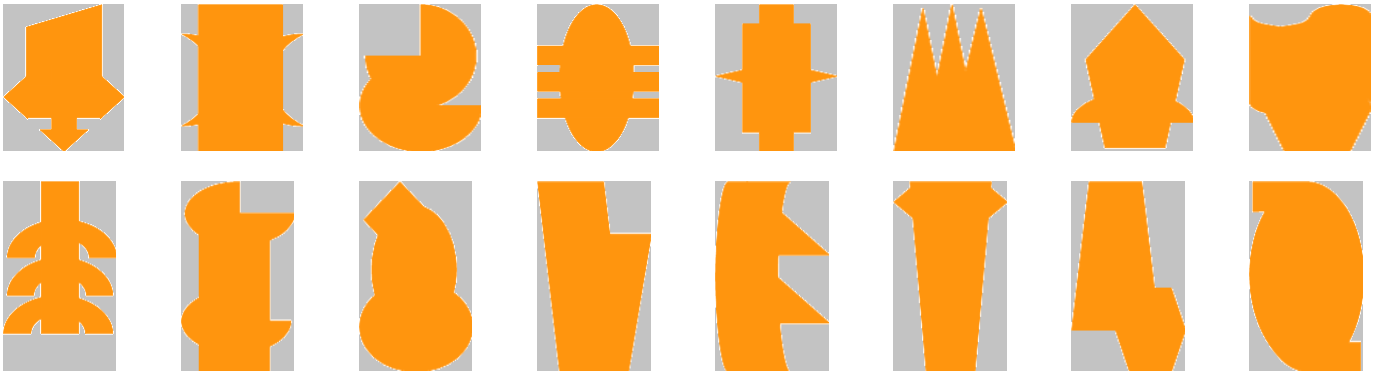


Figure 4a. Fractals used for non-linguistic material. Fractals in the upper row were used as fillers; fractals in the lower two rows were used in the triplets.



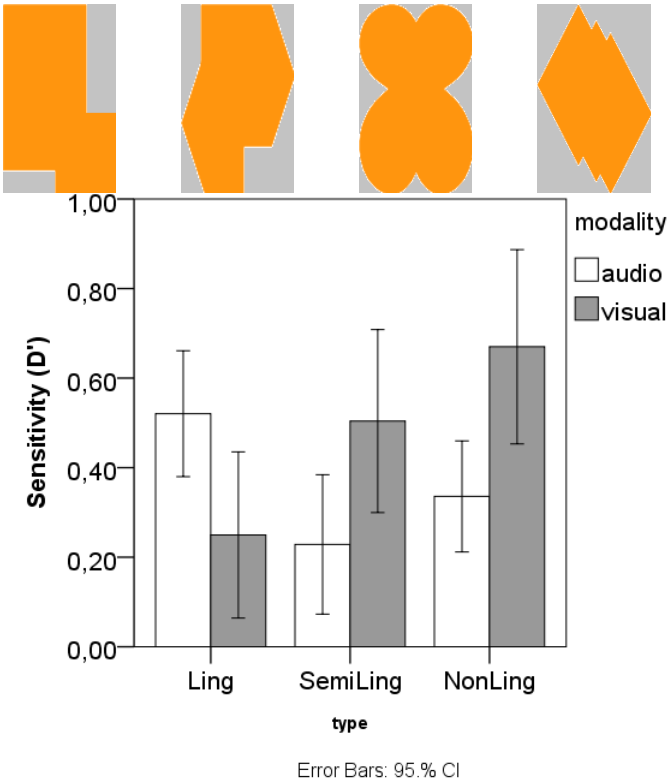


Figure 5a. Sensitivity to the embedded triplets across domains and modalities.

Figure 4b. Geometric shapes used for semi-linguistic material. Geometric shapes in the upper row were used as fillers; shapes in the lower two rows were used in the triplets.

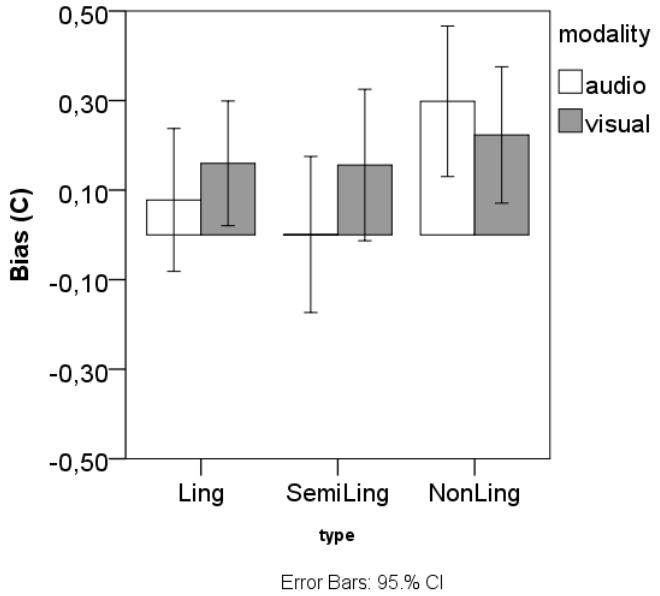


Figure 5b. Bias in regard to endorsing (positive) or rejecting (negative) test items as sequences from the learning input, across domains and modalities. A score of 0 indicates no bias.