eLife

Neuroscience

Statistical learning beyond words in human neonates

Reviewed Preprint
v2 • January 14, 2025
Revised by authors

Reviewed Preprint
v1 • October 29, 2024

Ana Fló, Lucas Benjamin, Marie Palu, Ghislaine Dehaene-Lambertz

Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France • Department of Developmental Psychology and Socialisation and Department of Neuroscience, University of Padova, Padova, Italy • Departement d'étude Cognitives, École Normale Supérieure, Paris, France • Aix Marseille Univ, INSERM, INS, Inst Neurosci syst, Marseille, France

eLife Assessment

The manuscript provides **important** new insights into the mechanisms of statistical learning in early human development, showing that statistical learning in neonates occurs robustly and is not limited to linguistic features but occurs across different domains. The evidence is **convincing** and the findings are highly relevant for researchers working in several domains, including developmental cognitive neuroscience, developmental psychology, linguistics, and speech pathology.

https://doi.org/10.7554/eLife.101802.2.sa4

Abstract

Interest in statistical learning in developmental studies stems from the observation that 8-month-olds were able to extract words from a monotone speech stream solely using the transition probabilities (TP) between syllables (Saffran et al., 1996). A simple mechanism was thus part of the human infant's toolbox for discovering regularities in language. Since this seminal study, observations on statistical learning capabilities have multiplied across domains and species, challenging the hypothesis of a dedicated mechanism for language acquisition. Here, we leverage the two dimensions conveyed by speech –speaker identity and phonemes– to examine (1) whether neonates can compute TPs on one dimension despite irrelevant variation on the other and (2) whether the linguistic dimension enjoys an advantage over the voice dimension. In two experiments, we exposed neonates to artificial speech streams constructed by concatenating syllables while recording EEG. The sequence had a statistical structure based either on the phonetic content, while the voices varied randomly (Experiment 1) or on voices with random phonetic content (Experiment 2). After familiarisation, neonates heard isolated duplets adhering, or not, to the structure they were familiarised with. In both experiments, we observed neural entrainment at the frequency of the regularity and distinct Event-Related Potentials (ERP) to correct and incorrect duplets, highlighting the universality of statistical learning mechanisms and suggesting it operates on virtually any dimension the input is factorised. However, only linguistic duplets elicited a specific ERP component, potentially an N400 precursor, suggesting a lexical stage triggered by phonetic regularities already at birth. These results show that, from birth, multiple input regularities can be processed in parallel and feed different higher-order networks.

Highlights

- Human neonates are sensitive to regularities in speech, encompassing both phonetic content and voice dimension.
- There is no observed advantage for statistical computations of linguistic content over voice content.
- Both speech dimensions are processed in parallel by distinct networks.
- Phonetic regularities evoked a specific ERP component in the post-learning phase, suggesting activations along the pathway to the lexicon.

Introduction

Since speech is a continuous signal, one of the infants' first challenges during language acquisition is to break it down into smaller units, notably to be able to extract words. Parsing has been shown to rely on prosodic cues (e.g., pitch and duration changes) but also on identifying regular patterns across perceptual units. Nearly 20 years ago, Saffran, Newport, and Aslin (1996) demonstrated that infants are sensitive to local regularities, specifically Transitional Probabilities (TP) between syllables—that is, the probability that two particular syllables follow each other in a given sequence, $TP = P(S_{i+1} | S_i)$. In their study, eight-month-old infants were exposed to a continuous, monotonous stream of syllables composed of four randomly concatenated tri-syllabic pseudo-words. Within each word, the sequence of syllables was fixed, resulting in a TP of 1 for each syllable pair within a word. By contrast, because each word could be followed by any of the three other words, the TP for syllable pairs spanning word boundaries dropped to 1/3. The authors found that, after only two minutes of exposure to this stream, infants could distinguish between the original "words" and "part-words"—syllable triplets, including the between-word TP drop. Since this seminal study, statistical learning has been regarded as an essential mechanism for language acquisition because it allows for the extraction of regular patterns without prior knowledge.

During the last two decades, many studies have extended this finding by demonstrating sensitivity to statistical regularities in sequences across domains and species. For example, segmentation capacities analogous to those observed for a syllable stream are observed throughout life in the auditory modality for tones (Kudo et al., 2011; Saffran et al., 1999) and in the visual domain for shapes (Bulf et al., 2011; Fiser & Aslin, 2002; Kirkham et al., 2002) and actions (Baldwin et al., 2008; Monroy et al., 2017). Non-human animals, including cotton-top tamarins (Hauser et al., 2001), rats (Toro & Trobalón, 2005), dogs (Boros et al., 2021) and chicks (Santolin et al., 2016), have also been shown to be sensitive to TPs between successive events. While the level of complexity that each species can track might differ, statistical learning appears as a general learning mechanism for auditory and visual sequence processing (for a review of statistical learning capacities across species, see Santolin and Saffran, 2018).

Using near-infra-red spectroscopy (NIRS) and electroencephalography (EEG), we have shown that statistical learning is observed in sleeping neonates (Fló et al., 2022; Fló et al., 2019), highlighting the automaticity of this mechanism. We also discovered that tracking statistical probabilities might not lead to stream segmentation in the case of quadrisyllabic words in both neonates and adults, revealing an unsuspected limitation of this mechanism (Benjamin et al., 2022). Here, we aimed to further characterise this mechanism to shed light on its role in the early stages of language acquisition. Specifically, we addressed two questions: First, we

investigated whether statistical learning in human neonates is a general learning mechanism applicable to any speech feature or whether there is a bias in favour of computations on linguistic content to extract words. Second, we explored the level at which newborns compute transitions between syllables, at a low auditory level (i.e. between the presented events) or at a later phonetic level, after normalisation through irrelevant dimensions such as voices.

To test this, we have taken advantage of the fact that syllables convey two important pieces of information for humans: what is being said and who is speaking, i.e. linguistic content and speaker's identity. While statistical learning can be helpful to word extraction, a statistical relationship between successive voices is not of obvious use and could even hinder word extraction if instances of a word uttered by a different speaker are considered independently. However, as auditory processing is organised along several hierarchical and parallel pathways integrating different spectro-temporal dimensions (Belin et al., 2000; DeWitt & Rauschecker, 2012; Norman-Haignere et al., 2015; Zatorre & Belin, 2001), statistical learning might be computed on one dimension independently of the variation of the other along the linguistic and the voice pathways in parallel. Numerous behavioural and brain imaging studies have revealed phonetic normalisation across speakers in infants (Dehaene-Lambertz & Pena, 2001; Gennari et al., 2021; Kuhl & Miller, 1982). By the second half of the first year, statistical learning has also been shown to occur even when different voices are used (Graf Estes & Lew-Williams, 2015) or when natural speech is presented, where syllable production can vary from instance to instance (Hay et al., 2011; Pelucchi et al., 2015). Therefore, we hypothesised that neonates would compute TPs between syllables even when each syllable is produced by a different speaker, relying on a normalisation process at the syllable or phonetic level. However, our predictions regarding TPs learning across different voices were more open-ended. Either statistical learning is universal and can be similarly computed over any feature comprising voices, or listening to a speech stream favours the processing of phonetic regularities over other non-linguistic dimensions of speech and thus hinders the possibility of computing regularities over voices.

To study these possibilities, we constructed artificial streams using six consonant-vowel (CV) syllables produced by six voices (Table S1 and S2), resulting in 36 possible tokens (6 syllables × 6 voices). To form the streams, tokens were combined either by imposing structure to their phonetic content (Experiment 1: Structure over Phonemes) or their voice content (Experiment 2: Structure over Voices), while the second dimension varied randomly (**Figure 1**). The structure consisted of the random concatenation of three duplets (i.e., two-syllable units) defined only by one of the two dimensions. For example, in Experiment 1, one duplet could be *petu* with each syllable uttered by a random voice each time they appear in the stream (e.g *pe* is produced by voice^{1} and *tu* by voice^{6} in one instance and in another instance *pe* is produced by voice^{3} and *tu* by voice^{2}). In contrast, in Experiment 2, one duplet could be the combination [voice^{1}-voice^{6}], each uttering randomly any of the syllables.

If infants at birth compute regularities based on a neural representation of the syllable as a whole, i.e. comprising both phonetic and voice content, this would require computing a 36 × 36 TPs matrix relating each token. Under this computation, TPs in the structured stream would alternate between 1/6 within words and 1/12 between words. We predicted infants would fail the task in both experiments as previous studies showing successful segmentation in infants commonly use higher within-word TPs (usually 1) and far fewer tokens (typically 4 to 12) (Saffran & Kirkham, 2018). By contrast, if speech input is processed along the two studied dimensions in distinct pathways, it enables the calculation of two independent TP matrices of 6×6 between the six voices in a voice pathway and between the six syllables in a phonetic pathway. These computations would result in TPs alternating between 1 and 1/2 for the informative feature while remining uniform at 1/5 for the uninformative feature, leading to stream segmentation based on the informative dimension.

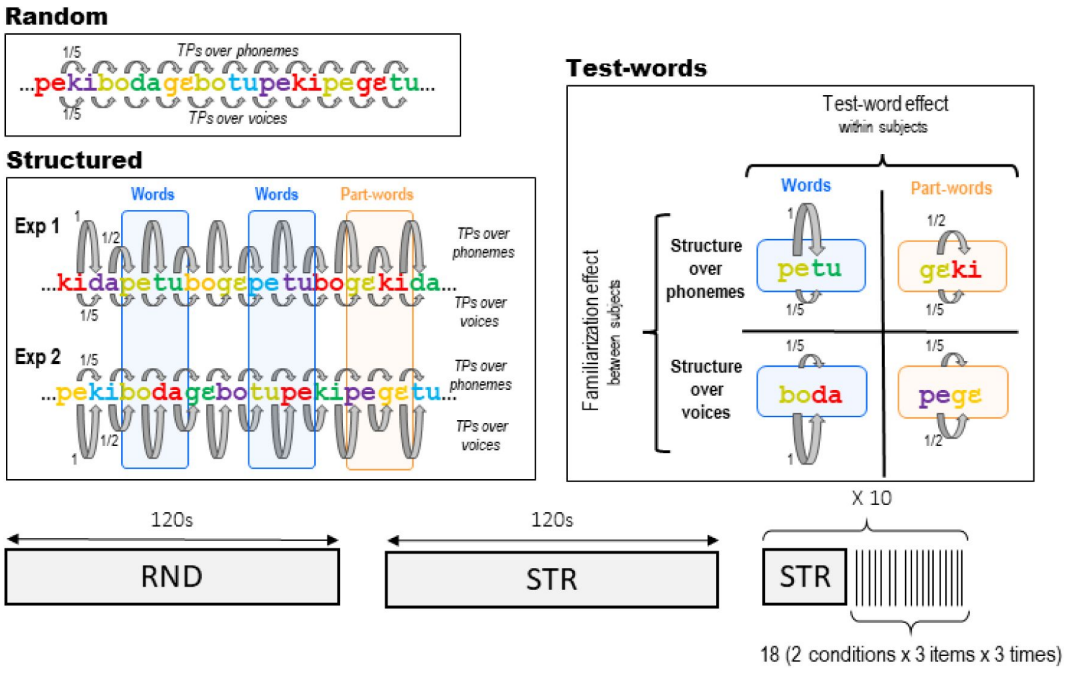


Figure 1.

Experimental protocol.

The experiments started with a Random stream (120 s) in which both syllables and voices changed randomly, followed by a long-Structured stream (120 s). Then, 10 short familiarisation streams (30 s), each followed by test blocks comprising 18 isolated duplets (SOA 2-2.3 s) were presented. Example streams are presented to illustrate the construction of the streams, with different colours representing different voices. In Experiment 1, the Structured stream had a statistical structure based on phonemes (TPs alternated between 1 and 0.5), while the voices were randomly changing (uniform TPs of 0.2). For example, the two syllables of the word *"petu"* were produced by different voices, which randomly changed at each presentation of the word (e.g. *"yellow"* voice and *"green"* voice for the first instance, *"blue"* and *"purple"* voice for the second instance, etc..). In Experiment 2, the statistical structure was based on voices (TPs alternated between 1 and 0.5), while the syllables changed randomly (uniform TPs of 0.2). For example, the *"green"* voice was always followed by the *"red"* voice, but they were randomly saying different syllables *"boda"* in the first instance, *"tupe"* in the second instance, etc... The test duplets in the recognition test phase were either Words (TP=1) or Partwords (TP=0.5). Words and Partwords were defined in terms of phonetic content for Experiment 1 and voice content for Experiment 2.

As in our previous experiments (Benjamin et al., 2022; Fló, Benjamin, et al., 2022), we used high-density EEG (128 electrodes) to study speech segmentation abilities. Using artificial language with syllables with a fixed duration elicits Steady State Evoked Potentials (SSEP) at the syllable rate. Crucially, if the artificial language presents a regular structure (i.e., regular drops in TPs marking word boundaries) and if the structure is perceived, then the neural responses reflect the slower frequency of the word as well (Buiatti et al., 2009). In other words, the brain activity becomes phase-locked to the regular input, increasing the Inter Trial Coherence (ITC) and power at the input regularity frequencies. Under these circumstances, the analysis in the frequency domain is advantageous since fast and periodic responses can be easily investigated by looking at the target frequencies without considering their specific timing (Kabdebon et al., 2022). The phenomenon is also named frequency tagging or neural entrainment in the literature. Here, we will refer to it indistinctively as SSEP or neural entrainment since we do not aim to make any hypothesis on the origin of the response (i.e., pure evoked response or phase reset of endogenous oscillations, Giraud and Poeppel, 2012).

Our study used an orthogonal design across two groups of 1-4 day-old neonates. In Experiment 1 (34 infants), the regularities in the speech stream were based on the phonetic content, while the voices varied randomly (Phoneme group). Conversely, in Experiment 2 (33 infants), regularities were based on voices, while the phonemes changed randomly (Voice group). Both experiments started with a control stream in which both features varied randomly (i.e. Random stream, 120 s). Next, neonates were exposed to the Structured stream (120 s) with statistical structure over one or the other feature. The experiments ended with ten sets of 18 test duplets presented in isolation, preceded by short Structured streams (30 s) to maintain learning (**Figure 1**). Half of the test duplets corresponded to familiar regularities (Words, TP =1), and the other half were duplets present in the stream but which straddled a drop in TPs (Partwords, TP = 0.5).

To investigate online learning, we quantified the ITC as a measure of neural entrainment at the syllable (4 Hz) and word rate (2 Hz) during the presentation of the continuous streams. For the recognition process, we compared ERPs to Word and Part-Word duplets. We also tested 57 adult participants in a comparable behavioural experiment to investigate adults' segmentation capacities under the same conditions.

Results

Neural entrainment during the familiarisation phase

To measure neural entrainment, we quantified the ITC in non-overlapping epochs of 7.5 s. We compared the studied frequency (syllabic rate 4 Hz or duplet rate 2 Hz) with the 12 adjacent frequency bins following the same methodology as in our previous studies.

For the Random streams, we observed significant entrainment at syllable rate (4Hz) over a broad set of electrodes in both experiments ($p < 0.05$, FDR corrected) and no enhanced activity at the duplet rate for any electrode ($p > 0.05$, FDR corrected). Concerning the Structured streams, ITC increased at both the syllable and duplet rate ($p < 0.05$, FDR corrected) in both experiments (**Figure 2A**, 2B). The duplet effect was localised over occipital and central-left electrodes in the Phoneme group and over occipital and temporal-right electrodes in the Voice group.

We also directly compared the ITC at both frequencies of interest between the Random and Structured conditions (**Figure 2C**). We found electrodes with significantly higher ITC at the duplet rate during Structured streams than Random streams in both experiments ($p < 0.05$, FDR corrected). We also found electrodes with higher entrainment at syllable rate during the Structured than Random streams in both experiments ($p < 0.05$, FDR corrected). This effect might result from stronger or more phase-locked responses to syllables when the input is structured.

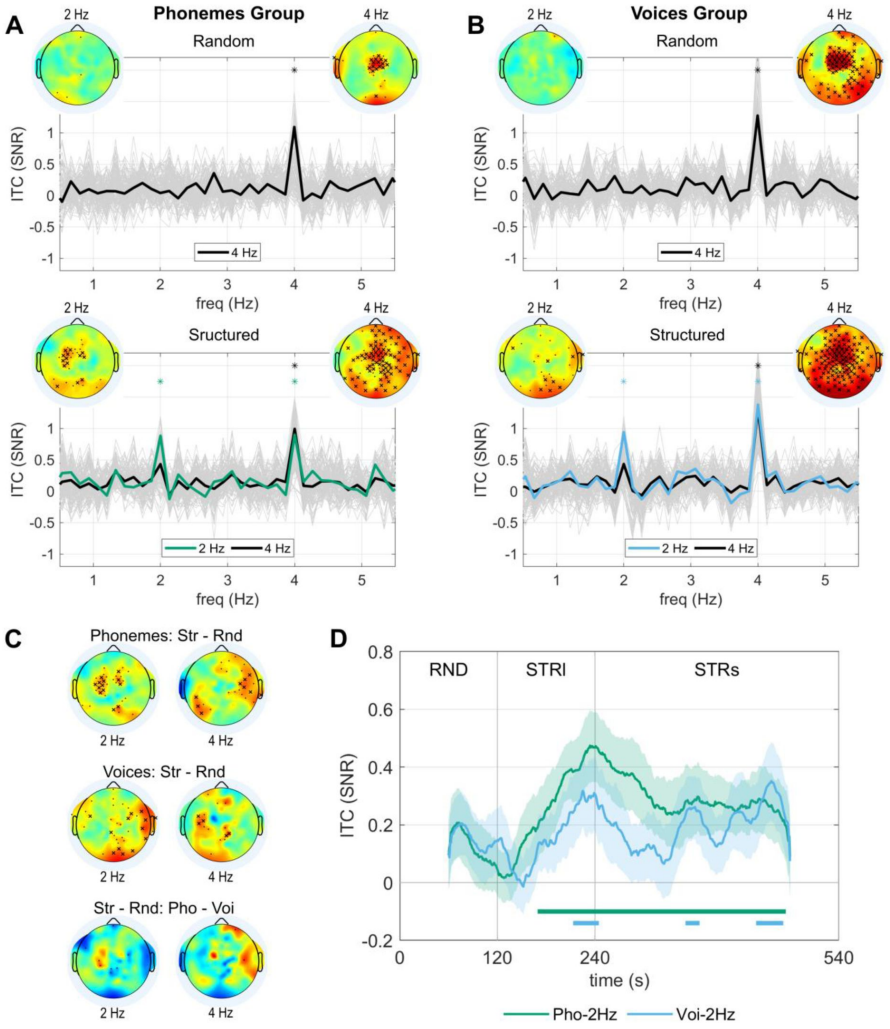


Figure 2.

Neural entrainment during the random and structured streams.

(A) SNR for the ITC during the Random and Structured streams of Experiment 1 (structure on phonetic content). The topographies represent the entrainment in the electrode space at the syllabic (4 Hz) and duplet rates (2 Hz). Crosses indicate the electrodes showing enhanced neural entrainment (cross: $p < 0.05$, one-sided paired permutation test, FDR corrected by the number of electrodes; dot: $p < 0.05$, without FDR correction). Colour scale limits [-1.8, 1.8]. The entrainment for each electrode is shown in light grey. The thick black line shows the mean over the electrodes with significant entrainment relative to the adjacent frequency bins at the syllabic rate (4 Hz) ($p < 0.05$ FDR corrected). The thick green line shows the mean over the electrodes showing significant entrainment relative to the adjacent frequency bins at the duplet rate (2 Hz) ($p < 0.05$ FDR corrected). The asterisks indicate frequency bins with entrainment significantly higher than on adjacent frequency bins for the average across electrodes ($p < 0.05$, one-sided permutation test, FDR corrected for the number of frequency bins). (B) Analog to A for Experiment 2 (structure on voice content). (C) The first two rows show the topographies for the difference in entrainment during the Structured and Random streams at 4 Hz and 2 Hz for both experiments. Crosses indicate the electrodes showing stronger entrainment during the Structured stream (cross: $p < 0.05$, one-sided paired permutation test, FDR corrected by the number of electrodes; dot: $p < 0.05$, without FDR correction). The bottom row shows the interaction effect by comparing the difference in entrainment during the Structured and Random streams between Experiments 1 and 2. Crosses indicate significant differences (cross: $p < 0.05$, two-sided unpaired permutation test, FDR corrected by the number of electrodes; dot: $p < 0.05$, without FDR correction). (D) Time course of the neural entrainment at 4 Hz for the average over electrodes showing significant entrainment during the Random stream and at 2 Hz for the average over electrodes showing significant entrainment during the Structured stream (Phoneme: green line, Voice blue line). The shaded area represents standard errors. The horizontal lines on the bottom indicate when the entrainment was larger than 0 ($p < 0.05$, one-sided t-test, corrected by FDR by the number of time points).

Since the first harmonic of the duplet rate (2 x 2 Hz) coincides with the syllable rate (4 Hz), word entrainment during the structured streams could also contribute to this effect. However, this contribution is unlikely since the electrodes showing higher 4 Hz entrainment during Structured than Random streams differ from those showing duplet-rate activity at 2 Hz.

Finally, we looked for an interaction effect between groups and conditions (Structured vs. Random streams) (**Figure 2C**). A few electrodes show differential responses between groups, reflecting the topographical differences observed in the previous analysis, notably the trend for stronger ITC at 2 Hz over the left central electrodes for the Phoneme group compared to the Voice group, but none survive multiple comparison corrections.

Learning time-course

To investigate the time course of the learning, we computed neural entrainment at the duplet rate in sliding time windows of 2 minutes with a 1 s step across both random and structured streams (**Figure 2D**). Notice that because the integration window was two minutes long, the entrainment during the first minute of the Structured stream included data from the random stream. To test whether ITC at 2 Hz increased during long Structured familiarisation (120 s), we fitted a Linear Mixed Model (LMM) with a fixed effect of time and random slopes and interceptions for individual subjects: $ITC \sim -1 + time + (1 + time|subject)$. In the Phoneme group, we found a significant time effect ($\beta=4.16\times10^{-3}$, 95% CI=[2.06×10^{-3}, 6.29×10^{-3}], SE=1.05×10^{-3}, p=4×10^{-4}), as well as in the Voice group ($\beta=2.46\times10^{-3}$, 95% CI=[2.6×10^{-4}, 4.66×10^{-3}], SE=1.09×10^{-3}, p=0.03). To test for differences in the time effect between groups, we included all data in a single LMM: $ITC \sim -1 + time * group + (1 + time|subject)$. The model showed a significant fixed effect of time for the Phoneme group consistent with the previous results ($\beta=4.22\times10^{-3}$, 95% CI=[1.07×10^{-3}, 7.37×10^{-3}], SE=1.58×10^{-3}, p=0.0096), while the fixed effect estimating the difference between the Phoneme and Voice groups was not significant ($\beta=-1.84\times10^{-3}$, 95% CI=[-6.29×10^{-3}, 2.61×10^{-3}], SE=2.24×10^{-3}, p=0.4).

ERPs during the test phase

To test the recognition process, we also measured ERP to isolated duplets afterwards. The average ERP to all conditions merged is shown in Figure S1. We investigated (1) the main effect of test duplets (Word vs. Part-word) across both experiments, (2) the main effect of familiarisation structure (Phoneme group vs. Voice group), and finally (3) the interaction between these two factors. We used non-parametric cluster-based permutation analyses (i.e. without a priori ROIs) (Oostenveld et al., 2011).

The difference between Word and Part-word consisted of a dipole with a median positivity and a left temporal negativity ranging from 400 to 1500 ms, with a maximum around 800-900 ms (**Fig 3**). Cluster-based permutations recovered two significant clusters around 500-1500 ms: a frontal-right positive cluster (p = 0.019) and a left temporal negative cluster (p = 0.0056). A difference between groups was observed consisting of a dipole that started with a right temporal positivity left temporo-occipital negativity around 300 ms and rotated anti-clockwise to bring the positivity over the frontal electrodes and the negativity at the back of the head (500-800 ms) (Fig S2). Cluster-based permutations on the Phoneme group vs. Voice group recovered a posterior cluster (p = 0.018) around 500 ms; with no positive cluster reaching significance (p >0.10). A cluster-based permutation analysis on the interaction effect, i.e., comparing Words - Part-Words between both experiments, showed no significant clusters (p > 0.1).

As cluster-based statistics are not very sensitive, we also analysed the ERPs over seven ROIS defined on the grand average ERP of all merged conditions (see Methods). Results replicated what we observed with the cluster-based permutation analysis with similar differences between Words and Part-words for the effect of familiarisation and no significant interactions. Results are presented in SI. The temporal progression of voltage topographies for all ERPs is presented in Figure S2. To verify that the effects were not driven by one group per duplet type condition, we

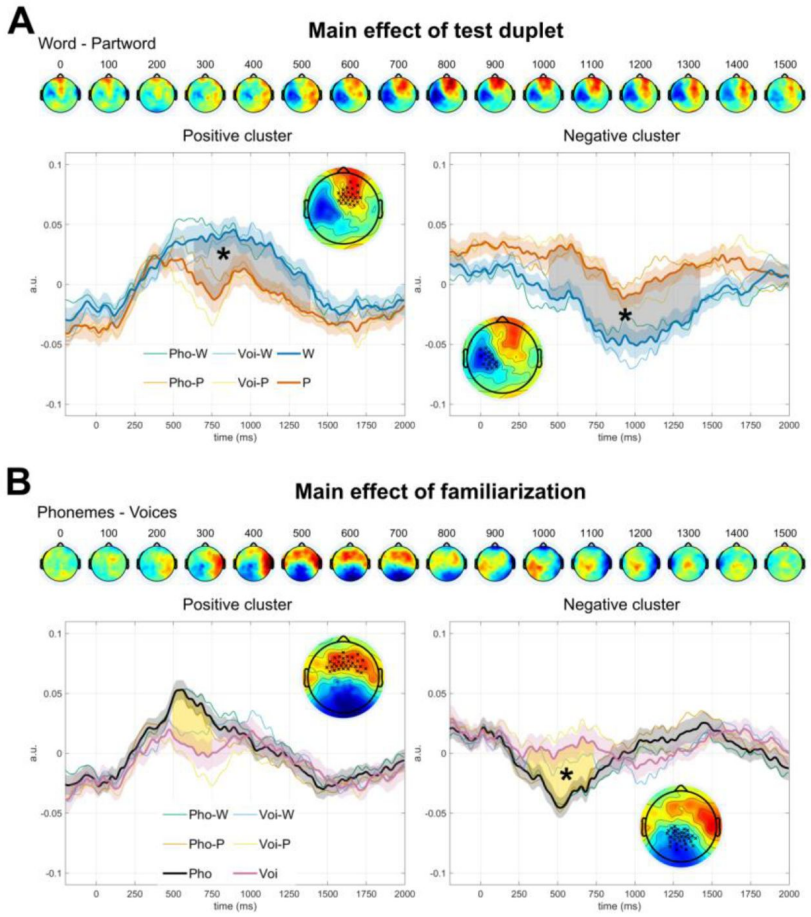


Figure 3.

Cluster-based permutation analysis of ERPs to isolated duplets during recognition

The topographies show the difference between the two conditions corresponding to each main effect. Results obtained from the cluster-based permutation analyses are shown at the bottom of each panel. Thick lines correspond to the grand averages for the two main tested conditions. Shaded areas correspond to the standard error across participants. Thin lines show the ERPs separated by duplet type and familiarisation type. The shaded areas between the thick lines show the time extension of the cluster. The topographies correspond to the difference between conditions during the time extension of the cluster. The electrodes belonging to the cluster are marked with a cross. Significant clusters are indicated with an asterisk. Color scale limits [-0.07, 0.07] a.u. (A) Main effect of Test-duplets (Words - Part-words) over a frontal-right positive cluster ($p = 0.019$) and a left temporal negative cluster ($p = 0.0056$). (B) Main effect of familiarisation (Phonemes - Voices) over a posterior negative cluster ($p = 0.018$). The frontal positive cluster did not reach significance ($p = 0.12$). Results are highly comparable to the ROIs-based analysis presented in SI (Fig S3).

ran a mixed two-way ANOVA for the average activity in each ROI and significant time window, with duplet type (Word/Part-word) as within-subjects factor and familiarisation as between-subjects factor. We did not observe significant interactions in any case. Future studies should consider a within-subject design to gain sensitivity to possible interaction effects.

Adult's behavioural performance in the same task

Adult participants heard a Structure Learning stream lasting 120 s and then ten sets of 18 test duplets preceded by Short Structure streams (30 s). For each test duplet, they had to rate its familiarity on a scale from 1 to 6. For the group familiarised with the Phoneme structure, there was a significant difference between the scores attributed to Words and Part-words (*t(26)=2.92, p = 0.007, Cohen's d =0.562*). The difference was marginally significant for the group familiarised with the Voice structure (*t(29)=2.0443, p = 0.050, Cohen's d=0.373*) (**Figure 4**). A 2-way ANOVA with test-duplets and familiarisation as factors revealed a main effect of Word (*F(1,55)=12.52, p=0.0008, η^2=0.039*), no effect of familiarisation (*F(1,55) <1*), and a significant interaction Word × Familiarisation (*F(1,55) = 5.28, p= 0.025, η_g^2=0.017*).

Discussion

Statistical learning is a general learning mechanism

In two experiments, we compared auditory statistical learning over a linguistic and a non-linguistic dimension in sleeping neonates. We took advantage of the possibility of constructing streams based on the same 36 tokens (6 syllables × 6 voices), the only difference between the experiments being the arrangement of the tokens in the streams. We showed that neonates were sensitive to regularities based either on the phonetic or the voice dimensions of speech, even in the presence of a non-informative feature that must be disregarded.

Parsing based on statistical information was revealed by steady-state evoked potentials at the duplet rate observed around 2 min after the onset of the familiarisation stream and by different ERPs to Words and Part-words presented during a recognition phase in both experiments. Despite variations in the other dimension, statistical learning was possible, showing that this mechanism operates at a stage when these dimensions have already been separated along different processing pathways. Our results, thus, revealed that linguistic content and voice identity are calculated independently and in parallel. This result confirms that, even in newborns, the syllable is not a holistic unit (Gennari et al., 2021) but that the rich temporo-frequential spectrum of speech is processed in parallel along different networks, probably using different integration factors (Boemio et al., 2005; Moerel et al., 2012; Zatorre & Belin, 2001). While statistical learning has already been described in many domains in neonates and young infants (Fló, Benjamin, et al., 2022; Fló et al., 2019; Kirkham et al., 2002; Kudo et al., 2011), we add here that even sleeping neonates distinctly applied it to potentially all dimensions in which speech is factorised in the auditory cortex (Gennari et al., 2021; Gwilliams et al., 2022). Our result contribute to a growing body of evidence supporting the universality of statistical learning —albeit with potential quantitative differences (Frost et al., 2015; Ren & Wang, 2023; Santolin & Saffran, 2018). This mechanism might be rooted in associative learning processes relying on the co-existence of event representations driven by slow activation decays (Benjamin et al., 2024).

We observed no clear processing advantage for the linguistic dimension over the voice dimension in neonates. The ability to track regularities in parallel for different speech features provides newborns with a powerful tool to create associations between recurring events and uncover structure. Future work could explore whether they can simultaneously track multiple regularities.

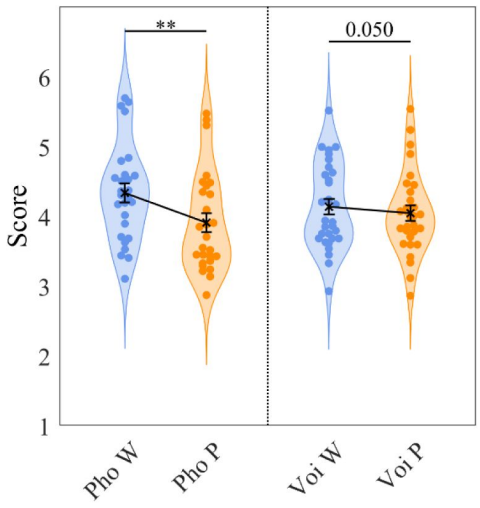


Figure 4.

Adults' behavioural experiment.

Each subject's average score attributed to the Words (blue) and Partwords (orange) is represented. On the right, for the group familiarised with the Phoneme structure and on the left, for the group familiarised with the Voice structure. The difference between test duplets was significant for the Phoneme group ($p = 0.007$) and only marginally significant for the Voice group ($p = 0.050$). There was also a significant interaction group × duplet type (p= 0.025).

Differences between statistical learning over voices and over phonemes

While the main pattern of results between experiments was comparable, we did observe some differences. The word-rate steady-state response (2 Hz) for the group of infants exposed to structure over phonemes was over posterior electrodes and left lateralised over central electrodes, while the group of infants hearing structure over voices showed mostly entrainment over posterior and right temporal electrodes. Auditory ERPs, after reference-averaged, typically consist of a central positivity and posterior negativity. These results are consistent with statistical learning in distinct lateralised neural networks to process speech's phonetic and voice content. Recent brain imaging studies in infants do indeed show hemispheric biases in precursors of later networks (Blasi et al., 2011; Dehaene-Lambertz et al., 2010; Mahmoudzadeh et al., 2013), even if specialisation increases during development (Shultz et al., 2014; Sylvester et al., 2023). However, the hemispheric differences reported here should be considered cautiously since the group comparison did not survive multiple comparison corrections. Future work investigating the neural networks involved should implement a within-subject design to gain statistical power.

The time course of entrainment at the duplet rate revealed that it emerged at a similar time for both statistical structures. While the duplet rate response seemed more stable in the Phoneme group (evidenced by a sustained ITC at the word rate above zero and a steeper slope of increase), no significant difference was observed between the groups. Furthermore, since both groups exhibited differences in the Words vs Part-words comparison during the recognition phase, it is unlikely that the differences observed during exposure to the stream were due to poorer computation of statistical transitions for voice regularities compared to phoneme regularities. Another explanation may lie not in the statistical calculation but in the stability of the voice representation itself, as voice processing challenges have been reported in both children and adults (Johnson et al., 2011; Mahmoudzadeh et al., 2016). This could disrupt stable chunking during the stream even while preserving word recognition during the test. In a previous study, we observed that entrainment failed to emerge under certain conditions despite neonates successfully computing TPs—likely due to the absence of chunking (Benjamin et al., 2022).

Phoneme regularities might trigger a lexical search

In the test phase using isolated duplets, we observed a significant difference between groups: A dipole, consisting of a posterior negative pole and a frontal positivity, was recorded around 500 ms following linguistic duplets but not voice duplets. Since the acoustic properties of the duplets were identical in both experiments, the difference can only be attributed to an endogenous process modulating duplet processing driven by the type of regularity to which neonates were exposed during the familiarisation phase.

Given its topography, latency and the phonetic context in which it appears, this component might be a precursor of the N400, a negative deflection at 200-600 ms over central-parietal electrodes elicited by lexico-semantic manipulations in adults (Kutas & Federmeier, 2011). Previous studies have already postulated components analogous to the N400 in infants. For example, a posterior negativity has been reported in infants as young as five months when hearing their own name compared to a stranger's name (Parise et al., 2010), when a pseudo-word was consistently vs inconsistently associated with an object (Friedrich & Friederici, 2011), and when they saw unexpected vs expected actions (Reid et al., 2009). These results suggest that such negativity might be related to semantic processing. As is often the case in infants, the latency of the component was delayed, and its topography was more posterior compared to older developmental stages (Friedrich & Friederici, 2005; Junge et al., 2021).

Furthermore, even in the absence of clear semantic content, an N400 has been reported in adults listening to artificial languages. For instance, Sanders et al. (2002) observed an N400 in adults listening to an artificial language after prior exposure to isolated pseudo-words. Other studies have demonstrated larger N400 amplitudes in adults when listening to structured streams compared to random sequences, whether these sequences consisted of syllables (Cunillera et al., 2006, 2009), tones (Abla et al., 2008), or shapes (Abla & Okanoya, 2009). Comparing ERPs from the recognition phase with those elicited by duplets during the familiarisation stream was not feasible due to the weaker signal-to-noise ratio in continuous stream recordings versus isolated stimuli and baseline challenges specific to infant background EEG (Eisermann et al., 2013). However, some inferences can still be drawn regarding the differences between the phoneme and voice experiments. Neural responses in both the learning and recognition phases could reflect a top-down effect: Phonetic regularities, but not voice regularities, would induce a lexical search during the presentation of the isolated duplets or at least activate a proto-lexical network.

Previous evidence suggests that infants extract and store possible word forms even before associating them with a clear meaning (Jusczyk & Hohne, 1997). For example, stronger fMRI activation for forward speech than backward speech in the left angular gyrus in 3-month-old infants has been linked to the activations of possible word forms in a proto-lexicon for native language sentences (Dehaene-Lambertz et al., 2002). Similarly, Shukla et al. (2011) demonstrated that chunks extracted from the speech stream serve as candidate words to which meanings can be attached. In their study, 6-month-olds spontaneously associated a pseudo-word extracted from natural sentences with a visual object. Although their experiment relied on TP and prosodic cues to extract the word, while our study only used statistical cues, this spontaneous bias to treat possible word forms as referring to a meaning (see also Bergelson and Aslin, 2017) might trigger activation along a lexicon pathway, explaining the difference between the two experiments: Only speech chunks based on phonetic regularities, and not voice regularities, are viable word candidates.

A lexical entry might also explain the more sustained activity during the familiarisation stream in the phonemes group, as the chunk might be encoded as a putative word in this admittedly rudimentary but present lexical store. In this hypothesis, the neural entrainment may reflect not only TPs but also the recovery of the “lexical” item.

Adults also learn voice regularities

Finally, we would like to emphasise that it is highly unnatural for a word not to be produced by the same speaker, nor for speakers to exhibit the kind of statistical relationships used here. Despite several years of exposure to speech, adults still demonstrated some learning of voice duplets, supporting the hypothesis of a general and automatic statistical learning ability. However, they also clearly displayed an advantage for phonetic regularities over voice regularities, revealed by the significant interaction Familiarisation × Word, not observed in neonates.

This difference might have several not mutually exclusive explanations. First, it may stem from the behavioural test being a more explicit measure of word recognition than the implicit task allowed by EEG recordings. Second, adults may perform better with phoneme structure due to a more effective auditory normalisation process or the additional use of a writing code for phonemes, which does not exist for voices. Finally, neonates, having little experience and therefore fewer expectations or constraints, might serve as better revealers of the possibilities afforded by statistical learning compared to older participants.

Conclusion

Altogether, our results show that statistical learning works similarly on different speech features in human neonates with no clear advantage for computing linguistically relevant regularities in speech. This supports the idea that statistical learning is a general learning mechanism, probably operating on common computational principles across neural networks (Benjamin et al., 2024) but within different neural networks with a different chain of operations: phonetic regularities induce a supplementary component –not seen in the case of voice regularities—that we related to a lexical N400. Understanding how statistical learning computations over linguistically relevant dimensions, such as the phonetic content of speech, are extracted and passed on to subsequent processing stages might be fundamental to uncovering how the infant brain acquires language. Further research is needed to understand how extracting regularities over different features articulates the language network.

Methods

Participants

Participants were healthy-full-term neonates with normal pregnancy and birth (GA > 38 weeks, Apgar scores ≥ 7/8 at 1/5 minute, birthweight > 2.5 Kg, cranial perimeter ≥ 33.0 cm), tested at the Port Royal Maternity (AP-HP), in Paris, France. Parents provided informed consent. The regional ethical committee for biomedical research (Comité de Protection des Personnes Region Centre Ouest 1, EudraCT/ID RCB: 2017-A00513-50) approved the protocol, and the study was carried out according to relevant guidelines and regulations. 67 participants (34 in Experiment 1 and 33 in Experiment 2) who provided enough data without motion artefacts were included (Experiment 1: 19 females; 1 to 4 days old; mean GA: 39.3 weeks; mean weight: 3387 g; Experiment 2: 15 females; 1 to 4 days old; mean GA: 39.0 weeks; mean weight: 3363 g). 12 other infants were excluded from the analyses (11 due to fussiness; 1 due to bad data quality).

Stimuli

The stimuli were synthesised using the MBROLA diphone database (Dutoit et al., 1996). Syllables had a consonant-vowel structure and lasted 250 ms (consonants 90 ms, vowels 160 ms). Six different syllables (*ki*, *da*, *pe*, *tu*, *bo*, *gɛ*) and six different voices were used (*fr3*, *fr1*, *fr7*, *fr2*, *it4*, *fr4*), resulting in a total of 36 syllable-voice combinations, from now on, tokens. The voices could be female or male and have three different pitch levels (low, middle, and high) (Table S1). We performed post-hoc tests to ensure that the results were not driven by a perception of two voices: female and male (see SI). The 36 tokens were synthesised independently in MBROLA, their intensity was normalised, and the first and last 5 ms were ramped to zero to avoid "clicks." The streams were synthesised by concatenating the tokens' audio files, and they were ramped up and down during the first and last 5 s to avoid the start and end of the stream serving as perceptual anchors.

The Structured streams were created by concatenating the tokens in such a way that they resulted in a semi-random concatenation of the duplets (i.e., pseudo-words) formed by one of the features (syllable/voice) while the other feature (voice/syllable) vary semi-randomly. In other words, in Experiment 1, the order of the tokens was such that Transitional Probabilities (TPs) between syllables alternated between 1 (within duplets) and 0.5 (between duplets), while between voices, TPs were uniformly 0.2. The design was orthogonal for the Structured streams of Experiment 2 (i.e., TPs between voices alternated between 1 and 0.5, while between syllables were evenly 0.2). The random streams were created by semi-randomly concatenating the 36 tokens to achieve

uniform TPs equal to 0.2 over both features. The semi-random concatenation implied that the same element could not appear twice in a row, and the same two elements could not repeatedly alternate more than two times (i.e., the sequence $X_kX_jX_kX_j$, where X_k and X_j are two elements, was forbidden). Notice that with an element, we refer to a duplet when it concerns the choice of the structured feature and to the identity of the second feature when it involves the other feature. The same statistical structures were used for both Experiments, only changing over which dimension the structure was applied. The learning stream lasted 120 seconds, with each duplet appearing 80 times. The 10 short structured streams lasted 30 seconds each, each duplet appearing a total of 200 times (10 × 20). The same random stream was used for both Experiments, and it lasted 120 seconds.

In Experiment 1, the duplets were created to prevent specific phonetic features from facilitating stream segmentation. In each experiment, two different structured streams (lists A and B) were used by modifying how the syllables/voices were combined to form the duplets (Table S2). Crucially, the Words/duplets of list A are the Part-words of list B and vice versa any difference between those two conditions can thus not be caused by acoustical differences. Participants were randomly assigned and balanced between lists and Experiments.

The test words were duplets formed by the concatenation of two tokens, such that they formed a Word or a Part-word according to the structured feature.

Procedure and data acquisition

Scalp electrophysiological activity was recorded using a 128-electrode net (Electrical Geodesics, Inc.) referred to the vertex with a sampling frequency of 250 Hz. Neonates were tested in a soundproof booth while sleeping or during quiet rest. The study involved: (1) 120 s of a random stream, (2) 120 s of a structured stream, (3) 10 series of 30 s of structured streams followed by 18 test sequences (SOA 2-2.3s).

Data pre-processing

Data were band-pass filtered 0.1-40 Hz and pre-processed using custom MATLAB scripts based on the EEGLAB toolbox 2021.0 according to the APICE pre-processing pipeline to recover as much free-artifacts data as possible (Fló, Gennari, et al., 2022).

Neural entrainment

The pre-processed data were further high-pass filtered at 0.2 Hz. Then, data was segmented from the beginning of each phase into 0.5 s long segments (240 duplets for the Random, 240 duplets for the long Structured, and 600 duplets for the short Structured). Segments containing samples with artefacts defined as bad data in more than 30% of the channels were rejected, and the remaining channels with artefacts were spatially interpolated.

Neural entrainment per condition

The 0.5 s epochs belonging to the same condition were reshaped into non-overlapping epochs (Benjamin et al., 2021) of 7.5 s (15 duplets, 30 syllables), retaining the chronological order; thus, the timing of the steady-state response. Subjects who did not provide at least 50% artifact-free epochs for each condition (at least 8 long epochs during Random and 28 during Structured) were excluded from the entrainment analysis (32 included subjects in Experiment 1, and 32 included subjects in Experiment 2). The retained subjects for Experiment 1, on average provided 13.59 epochs for the Random condition (SD 2.07, range [8, 16]) and 48.16 for the Structured conditions (SD 5.89, range [33, 55]). The retained subjects for Experiment 2, on average provided 13.78 epochs for the Random condition (SD 1.93, range [8, 16]) and 46.88 for the Structured conditions (SD 5.62, range [36, 55]). After data rejection, data were referenced to the average and normalized by dividing by the standard deviation within an epoch across electrodes and time. Next, data were

converted to the frequency domain using the Fast Fourier Transform (FFT) algorithm, and the ITC was estimated for each electrode during each condition (Random, Structured) as $ITC(f) = \frac{1}{N}\left|\sum_{i=1}^{N} e^{i\varphi(f,i)}\right|$, where N is the number of trials and ϕ(f,i) is the phase at frequency f and trial i. The ITC ranges from 0 to 1 (i.e., completely desynchronized activity to perfectly phased locked activity). Since we aim to detect an increase in signal synchronization at specific frequencies, the SNR was computed relative to the twelve adjacent frequency bins (six of each side corresponding to 0.8 Hz) (Kabdebon et al., 2022). This procedure also enables correcting differences in the ITC due to a different number of trials. Specifically, the SNR was $SNR(f) = (ITC(f)\text{-}mean(ITC_{noise}(f)))/std(ITC_{noise}(f))$, where $ITC_{noise}(f)$ is the ITC over the adjacent frequency bins. For statistical analysis, we compared the SNR at syllable rate (4 Hz) and duplet rate (2 Hz) against the average SNR over the 12 adjacent frequency bins using a one-tail paired permutation test (5000 permutations). We also directly compared the entrainment during the two conditions to individuate the electrodes showing a greater entrainment during the Structured than Random streams. We evaluated the interaction between Stream type (Random and Structured) and Familiarization type (Structured over Phonemes or Voices) by comparing the difference in entrainment between Structured and Random during the two experiments using a two sides unpaired permutation test (5000 permutations). All P-values were corrected across electrodes by FDR.

Neural entrainment time course

The 0.5 s epochs were concatenated chronologically (2 minutes of Random, 2 minutes of long Structured stream, and 5 minutes of short Structured blocks). The same analysis as above was performed in sliding time windows of 2 minutes with a 1 s step. A time window was considered valid if at least 8 out of the 16 epochs were free of motion artefacts. Missing values due to the presence of motion artifacts where linearly interpolated. Then, the entrainment time course at the syllable rate was computed as the average over the electrodes showing significant entrainment at 4 Hz during the Random condition, and at the duplet rate, as the average over the electrodes showing significant entrainment at 2 Hz during the Structured condition. Finally, data was smooth over a time window of 30 s.

To investigate the increase in the neural activity locked to the regularity during the long familiarisation, we fitted a LMM for each group of subjects. We included time as a fixed effect and random slopes and interceptions for individual subjects: $ITC \sim -1 + time + (1 + time|subject)$. We then compare the time effect between groups by including all data in a single LMM with time and group as fixed effects: $ITC \sim -1 + time * group\ (1 + time|subject)$.

ERPs to test words

The pre-processed data were filtered between 0.2 and 20 Hz, and epoched between [-0.2, 2.0] s from the onset of the duplets. Epochs containing samples identified as artifacts by APICE procedure were rejected. Subjects who did not provide at least half of the trials (45 trials) per condition were excluded (34 subjects kept for Experiment 1, and 33 for Experiment 2). No subject was excluded based on this criterion in the Phoneme groups, and one subject was excluded in the Voice groups. For Experiment 1, we retained on average 77.47 trials (SD 9.98, range [52, 89]) for the Word condition and 77.12 trials (SD 10.04, range [56, 89]) for the Part-word condition. For Experiment 2, we retained on average 73.73 trials (SD 10.57, range [47, 90]) for the Word condition and 74.18 trials (SD 11.15, range [46, 90]) for the Part-word condition. Data were reference averaged and normalised within each epoch by dividing by the standard deviation across electrodes and time.

Since the grand average response across both groups and conditions returned to the pre-stimulus level at around 1500 ms, we defined [0, 1500] ms as time windows of analysis. We first analysed the data using non-parametric cluster-based permutation analysis (Oostenveld et al., 2011) in the time window [0, 1500] ms (alpha threshold for clustering 0.10, neighbour distance ≤ 2.5 cm, clusters minimum size 3 and 5,000 permutations).

We also analysed the data in seven ROIs to ensure that no other effects were present that were not caught by the cluster-based permutation analysis. By inspecting the grand average ERP across both experiments and conditions, we identified three characteristic topographies: (a) positivity over central electrodes, (b) positivity over frontal electrodes and negativity over occipital electrodes, and (c) positivity over prefrontal electrodes and negativity over temporal electrodes (Figure S1). Then, we defined 7 symmetric ROIs: Central, Frontal Left, Frontal Right, Occipital, Prefrontal, Temporal Left, Temporal Right. We evaluated the main effect of Test-word type by comparing the EPRs between Words and Partwords (paired t-test) and the main effect of Familiarization type by comparing ERPs between Experiment 1 (structured over Phonemes) and Experiment 2 (structure over Voices) (unpaired t-test). All p-values were FDR corrected by the number of time points (n = 376) and ROIs (n = 7). To test for possible interaction effects, we compared the difference between Words and Partwords between the two groups. To verify that the main effects were not driven by one condition or group, we computed the average on each of the time windows where a main effect was identified considering both the Test-word type and Familiarization type factors, and we ran a two ways-ANOVA (Test-word type x Familiarization type). Results are presented in SI (Figure S3).

Adults' behavioural experiment

57 French-speaking adults were tested in an online experiment analogous to the infant study through the Prolific platform. All participants provided informed consent and received monetary compensation for their participation. The study was approved by the Ethical Research Committee of Paris Saclay University under the reference CER-Paris-Saclay-2019-063. The same stimuli as in the infants' experiment were used. Participants first heard 2 min of familiarisation with the Structured stream. Then, they completed ten sessions of re-familiarisation and testing. Each re-familiarization lasted 30 s, and in each test session, all 18 test words were presented. The structure could be either over the phonetic or the voice content, and two lists were used (see Table S2). Participants were randomly assigned to one of the groups and to one list. The Phoneme group included 27 participants, and the Voice group 30 participants. Before starting the experiment, subjects were instructed to pay attention to an invented language because later, they would have to answer if different sequences adhered to the structure of the language. During the test phase, subjects were asked to scale their familiarity with each test word by clicking with a cursor on a scale from 1 to 6.

Acknowledgements

We want to thank all the families who participated in the study. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 695710).

Additional information

Author contributions

A.F. and G.D.L. conceptualised the research; A.F., L.B. and M.P. performed the research; A.F. analysed the data; and A.F., L.B., and G.D.L. wrote the paper

Declaration of interests

The authors declare no competing interests.

Additional files

Supplemental Information

References

Abla D., Katahira K., Okanoya K (2008) **On-line assessment of statistical learning by event-related potentials** *Journal of Cognitive Neuroscience* **20**:952–964 https://doi.org/10.1162/jocn.2008.20058

Abla D., Okanoya K (2009) **Visual statistical learning of shape sequences: An ERP study** *Neuroscience Research* **64**:185–190 https://doi.org/10.1016/j.neures.2009.02.013

Baldwin D., Andersson A., Saffran J., Meyer M (2008) **Segmenting dynamic human action via statistical structure** *Cognition* **106**:1382–1407 https://doi.org/10.1016/j.cognition.2007.07.005

Belin P., Zatorre R. J., Lafaille P., Ahad P., Pike B (2000) **Voice-selective areas in human auditory cortex** *Nature* **403**:309–312

Benjamin L., Dehaene-Lambertz G., Fló A (2021) **Remarks on the analysis of steady-state responses: Spurious artifacts introduced by overlapping epochs** *Cortex* https://doi.org/10.1016/j.cortex.2021.05.023

Benjamin L., Fló A., Palu M., Naik S., Melloni L., Dehaene-Lambertz G (2022) **Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates** *Developmental Science* **26** https://doi.org/10.1111/desc.13300

Benjamin L., Sablé-Meyer M., Fló A., Dehaene-Lambertz G., Roumi F. A (2024) **Long-Horizon Associative Learning Explains Human Sensitivity to Statistical and Network Structures in Auditory Sequences** *Journal of Neuroscience* **44** https://doi.org/10.1523/JNEUROSCI.1369-23.2024

Bergelson E., Aslin R. N (2017) **Nature and origins of the lexicon in 6-mo-olds** *Proceedings of the National Academy of Sciences* **114**:12916–12921 https://doi.org/10.1073/pnas.1712966114

Blasi A. *et al.* (2011) **Early Specialization for Voice and Emotion Processing in the Infant Brain** *Current Biology* **21**:1220–1224 https://doi.org/10.1016/j.cub.2011.06.009

Boemio A., Fromm S., Braun A., Poeppel D (2005) **Hierarchical and asymmetric temporal sensitivity in human auditory cortices** *Nature Neuroscience* **8**:389–395 https://doi.org/10.1038/nn1409

Boros M., Magyari L., Török D., Bozsik A., Deme A., Andics A (2021) **Neural processes underlying statistical learning for speech segmentation in dogs** *Current Biology* https://doi.org/10.1016/j.cub.2021.10.017

Buiatti M., Peña M., Dehaene-Lambertz G (2009) **Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses** *NeuroImage* **44**:509–519 https://doi.org/10.1016/j.neuroimage.2008.09.015

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

Cunillera T., Càmara E., Toro J. M., Marco-Pallares J., Sebastián-Galles N., Ortiz H., Pujol J., Rodríguez-Fornells A (2009) **Time course and functional neuroanatomy of speech segmentation in adults** *NeuroImage* **48**:541–553 https://doi.org/10.1016/j.neuroimage.2009.06.069

Cunillera T., Toro J. M., Sebasti??n-Gall??s N., Rodr??guez-Fornells A. (2006) **The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study** *Brain Research* **1123**:168–178 https://doi.org/10.1016/j.brainres.2006.09.046

Dehaene-Lambertz G., Dehaene S., Hertz-Pannier L (2002) **Functional Neuroimaging of Speech Perception in Infants** *Science* **298**:2013–2015 https://doi.org/10.1126/science.1077066

Dehaene-Lambertz G., Montavont A., Jobert A., Allirol L., Dubois J., Hertz-Pannier L., Dehaene S (2010) **Language or music, mother or Mozart? Structural and environmental influences on infants' language networks** *Brain and Language* **114**:53–65 https://doi.org/10.1016/j.bandl.2009.09.003

Dehaene-Lambertz G., Pena M (2001) **Electrophysiological evidence for automatic phonetic processing in neonates** *Neuroreport* **12**:3155–3158 https://doi.org/10.1097/00001756-200110080-00034

Delorme A., Makeig S (2004) **EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis** *Journal of Neuroscience Methods* **134**:9–21 https://doi.org/10.1016/j.jneumeth.2003.10.009

DeWitt I., Rauschecker J. P (2012) **Phoneme and word recognition in the auditory ventral stream** *Proceedings of the National Academy of Sciences of the United States of America* **109**:E505–514 https://doi.org/10.1073/pnas.1113427109

Dutoit T., Pagel V., Pierret N., Bataille F., van der Vrecken O. (1996) **The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes** *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96* **3**:1393–1396 https://doi.org/10.1109/ICSLP.1996.607874

Eisermann M., Kaminska A., Moutard M.-L., Soufflet C., Plouin P (2013) **Normal EEG in childhood: From neonates to adolescents** *Neurophysiologie Clinique/Clinical Neurophysiology* **43**:35–65 https://doi.org/10.1016/j.neucli.2012.09.091

Fiser J., Aslin R. N (2002) **Statistical learning of new visual feature combinations by infants** *Proceedings of the National Academy of Sciences of the United States of America* **99**:15822–15826 https://doi.org/10.1073/pnas.232472899

Fló A., Benjamin L., Palu M., Dehaene-Lambertz G (2022) **Sleeping neonates track transitional probabilities in speech but only retain the first syllable of words** *Scientific Reports* **12** https://doi.org/10.1038/s41598-022-08411-w

Fló A., Brusini P., Macagno F., Nespor M., Mehler J., Ferry A. L (2019) **Newborns are sensitive to multiple cues for word segmentation in continuous speech** *Developmental Science* **22** https://doi.org/10.1111/desc.12802

Fló A., Gennari G., Benjamin L., Dehane-Lambertz G (2022) **Automated Pipeline for Infants Continuous EEG (APICE): A flexible pipeline for developmental cognitive studies** *Developmental Cognitive Neuroscience* https://doi.org/10.1016/j.dcn.2022.101077

Friedrich M., Friederici A. D (2005) **Semantic sentence processing reflected in the event-related potentials of one- and two-year-old children** *NeuroReport* **16** https://doi.org/10.1097/01.wnr.0000185013.98821.62

Friedrich M., Friederici A. D (2011) **Word Learning in 6-Month-Olds: Fast Encoding– Weak Retention** *Journal of Cognitive Neuroscience* **23**:3228–3240 https://doi.org/10.1162/jocn_a_00002

Frost R., Armstrong B. C., Siegelman N., Christiansen M. H (2015) **Domain generality versus modality specificity: The paradox of statistical learning** *Trends in Cognitive Sciences* **19**:117–125 https://doi.org/10.1016/j.tics.2014.12.010

Gennari G., Marti S., Palu M., Flo A., Dehaene-Lambertz G (2021) **Orthogonal neural codes for speech in the infant brain** *Proceedings of the National Academy of Sciences* **118** https://doi.org/10.1073/pnas.2020410118

Giraud A.-L., Poeppel D (2012) **Cortical oscillations and speech processing: Emerging computational principles and operations** *Nature Neuroscience* **15**:511–517 https://doi.org/10.1038/nn.3063

Estes K. G., Lew-Williams C (2015) **Listening through voices: Infant statistical word segmentation across multiple speakers** *Developmental Psychology* **51**:1517–1528 https://doi.org/10.1037/a0039725

Gwilliams L., King J.-R., Marantz A., Poeppel D (2022) **Neural dynamics of phoneme sequences reveal position-invariant code for content and order** *Nature Communications* **13** https://doi.org/10.1038/s41467-022-34326-1

Hauser M. D., Newport E. L., Aslin R. N (2001) **Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins** *Cognition* **78**:53–64 https://doi.org/10.1016/S0010-0277(00)00132-3

Hay J. F., Pelucchi B., Estes K. G., Saffran J. R (2011) **Linking sounds to meanings: Infant statistical learning in a natural language** *Cognitive Psychology* **63**:93–106 https://doi.org/10.1016/j.cogpsych.2011.06.002

Johnson E. K., Westrek E., Nazzi T., Cutler A (2011) **Infant ability to tell voices apart rests on language experience: Infant voice discernment** *Developmental Science* **14**:1002–1011 https://doi.org/10.1111/j.1467-7687.2011.01052.x

Junge C., Boumeester M., Mills D. L., Paul M., Cosper S. H (2021) **Development of the N400 for Word Learning in the First 2 Years of Life: A Systematic Review** *Frontiers in Psychology* **12** https://doi.org/10.3389/fpsyg.2021.689534

Jusczyk P. W., Hohne E. A (1997) **Infants' Memory for Spoken Words** *Science* **277**:1984–1986 https://doi.org/10.1126/science.277.5334.1984

Kabdebon C., Flo A., de Heering A., Aslin R. (2022) **The power of rhythms: How steady-state evoked responses reveal early neurocognitive development** *NeuroImage* **119150** https://doi.org/10.1016/j.neuroimage.2022.119150

Kirkham N. Z., Slemmer J. A, Johnson S. P. (2002) **Visual statistical learning in infancy: Evidence for a domain-general learning mechanism** *Cognition* **83**:4–5

Kudo N., Nonaka Y., Mizuno N., Mizuno K., Okanoya K (2011) **On-line statistical segmentation of a non-speech auditory stream in neonates as demonstrated by event-related brain potentials** *Developmental Science* **14**:1100–1106 https://doi.org/10.1111/j.1467-7687.2011.01056.x

Kuhl P. K., Miller J. D (1982) **Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants** *Perception & Psychophysics* **31**:279–292 https://doi.org/10.3758/BF03202536

Kutas M., Federmeier K. D (2011) **Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP)** *Annual Review of Psychology* **62**:621–647 https://doi.org/10.1146/annurev.psych.093008.131123

Mahmoudzadeh M., Dehaene-Lambertz G., Fournier M., Kongolo G., Goudjil S., Dubois J., Grebe R., Wallois F (2013) **Syllabic discrimination in premature human infants prior to complete formation of cortical layers** *Proceedings of the National Academy of Sciences* **110**:4846–4851 https://doi.org/10.1073/pnas.1212220110

Mahmoudzadeh M., Wallois F., Kongolo G., Goudjil S., Dehaene-Lambertz G (2016) **Functional Maps at the Onset of Auditory Inputs in Very Early Preterm Human Neonates** *Cerebral Cortex* **27**:2500–2512 https://doi.org/10.1093/cercor/bhw103

Moerel M., De Martino F., Formisano E. (2012) **Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity** *Journal of Neuroscience* **32**:14205–14216 https://doi.org/10.1523/JNEUROSCI.1388-12.2012

Monroy C. D., Gerson S. A., Hunnius S (2017) **Toddlers' action prediction: Statistical learning of continuous action sequences** *Journal of Experimental Child Psychology* **157**:14–28 https://doi.org/10.1016/j.jecp.2016.12.004

Norman-Haignere S., Kanwisher N. G., McDermott J. H (2015) **Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition** *Neuron* **88**:1281–1296 https://doi.org/10.1016/j.neuron.2015.11.035

Oostenveld R., Fries P., Maris E., Schoffelen J.-M (2011) **FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data** *Computational Intelligence and Neuroscience* **2011**:1–9 https://doi.org/10.1155/2011/156869

Parise E., Friederici A. D., Striano T (2010) **"Did You Call Me?" 5-Month-Old Infants Own Name Guides Their Attention** *PLoS ONE* **5** https://doi.org/10.1371/journal.pone.0014208

Pelucchi B., Hay J. F., Saffran J. R (2015) **Statistical learning in a natural language by 8-month-old infants** *Child Development* **80**:674–685 https://doi.org/10.1111/j.1467-8624.2009.01290.x

Reid V. M., Hoehl S., Grigutsch M., Groendahl A., Parise E., Striano T (2009) **The neural correlates of infant and adult goal prediction: Evidence for semantic processing systems** *Developmental Psychology* **45**:620–629 https://doi.org/10.1037/a0015209

Ren J., Wang M (2023) **Development of statistical learning ability across modalities, domains, and languages** *Journal of Experimental Child Psychology* **226** https://doi.org/10.1016/j.jecp.2022.105570

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

Saffran J. R., Johnson E. K., Aslin R. N., Newport E. L (1999) **Statistical learning of tone sequences by human infants and adults** *Cognition* **70**:27–52 https://doi.org/10.1016/S0010-0277(98)00075-4

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

Santolin C., Rosa-Salva O., Vallortigara G., Regolin L (2016) **Unsupervised statistical learning in newly hatched chicks** *Current Biology* **26**:R1218–R1220 https://doi.org/10.1016/j.cub.2016.10.011

Santolin C., Saffran J. R (2018) **Constraints on Statistical Learning Across Species** *Trends in Cognitive Sciences* **22**:52–63 https://doi.org/10.1016/j.tics.2017.10.003

Shukla M., White K. S., Aslin R. N (2011) **Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants** *Proceedings of the National Academy of Sciences* **108**:6038–6043 https://doi.org/10.1073/pnas.1017617108

Shultz S., Vouloumanos A., Bennett R. H., Pelphrey K (2014) **Neural specialization for speech in the first months of life** *Developmental Science* **17**:766–774 https://doi.org/10.1111/desc.12151

Sylvester C. M. *et al.* (2023) **Network-specific selectivity of functional connections in the neonatal brain** *Cerebral Cortex* **33**:2200–2214 https://doi.org/10.1093/cercor/bhac202

Toro J. M., Trobalón J. B (2005) **Statistical computations over a speech stream in a rodent** *Perception & Psychophysics* **67**:867–875 https://doi.org/10.3758/BF03193539

Zatorre R. J., Belin P (2001) **Spectral and Temporal Processing in Human Auditory Cortex** *Cerebral Cortex* **11**:946–953

Author information

Ana Fló

Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France, Department of Developmental Psychology and Socialisation and Department of Neuroscience, University of Padova, Padova, Italy
ORCID iD: 0000-0002-3260-0559

For correspondence: ana.flo@unipd.it

Lucas Benjamin

Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France, Departement d'étude Cognitives, École Normale Supérieure, Paris, France, Aix Marseille Univ, INSERM, INS, Inst Neurosci syst, Marseille, France
ORCID iD: 0000-0002-9578-6039

Marie Palu
Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France

Ghislaine Dehaene-Lambertz
Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France
ORCID iD: 0000-0003-2221-9081

Editors

Reviewing Editor
Björn Herrmann
Baycrest Hospital, Toronto, Canada

Senior Editor
Barbara Shinn-Cunningham
Carnegie Mellon University, Pittsburgh, United States of America

Reviewer #1 (Public review):

Summary:

Parsing speech into meaningful linguistic units is a fundamental yet challenging task that infants face while acquiring the native language. Computing transitional probabilities (TPs) between syllables is a segmentation cue well-attested since birth. In this research, the authors examine whether newborns compute TPs over any available speech feature (linguistic and non-linguistic), or whether by contrast newborns favor computation of TPs over linguistic content over non-linguistic speech features such as speaker voice. Using EEG and the artificial language learning paradigm, they record the neural responses of two groups of newborns presented with speech streams in which either phonetic content or speaker voice are structured to provide TPs informative of word boundaries, while the other dimension provides uninformative information. They compare newborns' neural responses to these structured streams to their processing of a stream in which both dimensions vary randomly. After the random and structured familiarization streams, the newborns are presented with (pseudo)words as defined by their informative TPs, as well as partwords (that is, sequences that straddle a word boundary), extracted from the same streams. Analysis of the neural responses show that while newborns neural activity entrained to the syllabic rate (2 Hz) when listening to the random and structured streams, it additionally entrained at the word rate (4 Hz) only when listening to the structured streams, finding no differential response between the streams structured around voice or phonetic information. Newborns showed also different neural activity in response to the words and part words. In sum, the study reveals that newborns compute TPs over linguistic and non-linguistic features of speech, these are calculated independently, and linguistic features do not lead to a processing advantage.

Strengths:

This interesting research furthers our knowledge of the scope of the statistical learning mechanism, which is confirmed to be a general-purpose powerful tool that allows humans to extract patterns of co-occurring events while revealing no apparent preferential processing for linguistic features. To answer its question, the study combines a highly replicated and well-established paradigm, i.e. the use of an artificial language in which pseudowords are concatenated to yield informative TPs to word boundaries, with a state-of-the-art EEG

analysis, i.e. neural entrainment. The sample size of the groups is sufficient to ensure power, and the design and analysis are solid and have been successfully employed before.

Weaknesses:

There are no significant weaknesses to signal in the manuscript. However, in order to fully conclude that there is no obvious advantage for the linguistic dimension in neonates, future studies should pit both dimensions against each other, to determine whether statistical learning weighs linguistic and non-linguistic features equally, or whether phonetic content is preferentially processed.

To sum up, the authors achieved their central aim of determining whether TPs are computed over both linguistic and non-linguistic features, and their conclusions are supported by the results. This research is important for researchers working on language and cognitive development, and language processing, as well as for those working on cross-species comparative approaches.

Comments on revisions:

The authors have addressed my suggestions. I have no further comments.

https://doi.org/10.7554/eLife.101802.2.sa3

Reviewer #2 (Public review):

Summary:

The manuscript investigates to what degree neonates show evidence for statistical learning from regularities in streams of syllables, either with respect to phonemes or with respect to speaker identity. Using EEG, the authors found evidence for both, stronger entrainment to regularities as well as ERP differences in response to violations of previously introduced regularities. In addition, violations of phoneme regularities elicited an ERP pattern which the authors argue might index a precursor of the N400 response in older children and adults.

Strengths:

All in all, this is a very convincing paper, which uses a clever manipulation of syllable streams to target the processing of different features. The combination of neural entrainment and ERP analysis allows for the assessment of different processing stages, and implementing this paradigm in a comparably large sample of neonates is impressive.

Weaknesses

The authors addressed all the concerns I previously raised well and I have no further comments.

https://doi.org/10.7554/eLife.101802.2.sa2

Reviewer #3 (Public review):

Summary:

This study is focused on testing whether statistical learning (a mechanism for parsing the speech signal into smaller chunks) preferentially operates over certain features of the speech at birth in humans. The features under investigation are phonetic content and speaker identity. Newborns are tested in an EEG paradigm in which they are exposed to a long stream

of syllables. In Experiment 1, newborns are familiarized with a sound stream that comprises regularities (transitional probabilities) over syllables (e.g., "pe" followed by "tu" in "petu" with 1.0 probability) while the voices uttering the syllables remain random. In Experiment 2, newborns are familiarized with the same sound stream but, this time, the regularities are built over voices (e.g., "green voice" followed by "red voice" with 1.0 probability) while the concatenation of syllables stays random. At the test, all newborns listened to duplets (individual chunks) that either matched or violated the structure of the familiarization. In both experiments, newborns showed neural entrainment to the regularities implemented in the stream, but only the duplets defined by transitional probabilities over syllables (aka word forms) elicited a N400 ERP component. These results suggest that statistical learning operates in parallel and independently on different dimensions of the speech already at birth and that there seems to be an advantage for processing statistics defining word forms rather than voice patterns.

Strengths:

This paper presents an original experimental design that combines two types of statistical regularities in a speech input. The design is robust and appropriate for EEG with newborns. I appreciated the clarity of the Methods section. There is also a behavioral experiment with adults that acts like a control study for newborns. The research question is interesting, and the results add new information about how statistical learning works at the beginning of postnatal life, and on which features of the speech. The figures are clear and helpful in understanding the methods, especially the stimuli and how the regularities were implemented.

Weaknesses:

I appreciated how the authors addressed my previous comments and concerns. I am satisfied with the changes made by the authors. I believe the paper reads much better. Also, the adjustment to the theoretical framework suits well.

https://doi.org/10.7554/eLife.101802.2.sa1

Author response:

The following is the authors' response to the original reviews.

We thank the three reviewers for their positive comments and useful suggestions. We have implemented most of the reviewers' recommendations and hope the manuscript is clearer now.

The main modifications are:

- A revision of the introduction to better explain what Transitional Probabilities are and clarify the rationale of the experimental design

- A revision of the discussion

- To tune down and better explain the interpretation of the different responses between duplets after a stream with phonetic or voice regularities (possibly an N400).

- To better clarify the framing of statistical learning as a universal learning mechanism that might share computational principles across features (or domains).

Below, we provide detailed answers to each reviewer's point.

> ***Response to Reviewer 1:***
>
> *There are no significant weaknesses to signal in the manuscript. However, in order to fully conclude that there is no obvious advantage for the linguistic dimension in neonates, it would have been most useful to test a third condition in which the two dimensions were pitted against each other, that is, in which they provide conflicting information as to the boundaries of the words comprised in the artificial language.*
>
> *This last condition would have allowed us to determine whether statistical learning weighs linguistic and non-linguistic features equally, or whether phonetic content is preferentially processed.*

We appreciate the reviewers' suggestion that a stream with conflicting information would provide valuable insights. In the present study, we started with a simpler case involving two orthogonal features (i.e., phonemes and voices), with one feature being informative and the other uninformative, and we found similar learning capacities for both. Future work should explore whether infants—and humans more broadly—can simultaneously track regularities in multiple speech features. However, creating a stream with two conflicting statistical structures is challenging. To use neural entrainment, the two features must lead to segmentation at different chunk sizes so that their effects lead to changes in power/PLV at different frequencies—for instance, using duplets for the voice dimension and triplets for the linguistic dimension (or vice versa). Consequently, the two dimensions would not be directly comparable within the same participant in terms of the number of distinguishable syllables/voices, memory demand, or SNR given the 1/F decrease in amplitude of background EEG activity. This would involve comparisons between two distinct groups counter-balancing chunk size and linguistic non-linguistic dimension. Considering the test phase, words for one dimension would have been part-words for the other dimension. As we are measuring differences and not preferences, interpreting the results would also have been difficult. Additionally, it may be difficult to find a sufficient number of clearly discriminable voices for such a design (triplets imply 12 voices). Therefore, an entirely different experimental paradigm would need to be developed.

If such a design were tested, one possibility is that the regularities for the two dimensions are calculated in parallel, in line with the idea that the calculation of statistical regularities is a ubiquitous implicit mechanism (see Benjamin et al., 2024, for a proposed neural mechanism). Yet, similar to our present study, possibly only phonetic features would be used as word candidates. Another possibility is that only one informative feature would be explicitly processed at a time due to the serial nature of perceptual awareness, which may prioritise one feature over the other.

We added one sentence in the discussion stating that more research is needed to understand whether infants can track both regularities simultaneously (p.13, l.270 “Future work could explore whether they can simultaneously track multiple regularities.”).

Note: The reviewer’s summary contains a typo: syllabic rate (4 Hz) –not 2 Hz, and word rate (2 Hz) –not 4 Hz.

> ***Response to Reviewer 2:***
>
> *N400: I am skeptical regarding the interpretation of the phoneme-specific ERP effect as a precursor of the N400 and would suggest toning it down. While the authors are correct in that infant ERP components are typically slower and more posterior compared to adult components, and the observed pattern is hence consistent with an adult N400, at the same time, it could also be a lot of other things. On a functional level, I can't follow the author's argument as to why a violation in phoneme regularity should elicit an N400,*

> *since there is no evidence for any semantic processing involved. In sum, I think there is just not enough evidence from the present paradigm to confidently call it an N400.*

The reviewer is correct that we cannot definitively determine the type of processing reflected by the ERP component that appears when neonates hear a duplet after exposure to a stream with phonetic regularities. We interpreted this component as a precursor to the N400, based on prior findings in speech segmentation tasks without semantic content, where a ~400 ms component emerged when adult participants recognised pseudowords (Sander et al., 2002) or during structured streams of syllables (Cunillera et al., 2006, 2009). Additionally, the component we observed had a similar topography and timing to those labelled as N400 in infant studies, where semantic processing was involved (Parise et al., 2010; Friedrich & Friederici, 2011).

Given our experimental design, the difference we observed must be related to the type of regularity during familiarisation (either phonemes or voices). Thus, we interpreted this component as reflecting lexical search— a process which could be triggered by a linguistic structure but which would not be relevant to a non-linguistic regularity such as voices. However, we are open to alternative interpretations. In any case, this difference between the two streams reveals that computing regularities based on phonemes versus voices does not lead to the same processes.

We revised the abstract (p.2, l.33) and the discussion of this result (p.15, l.299), toning them down. We hope the rationale of the interpretation is clearer now, as is the fact that it is just one possible interpretation of the results.

> *Female and male voices: Why did the authors choose to include male and female voices? While using both female and male stimuli of course leads to a higher generalizability, it also introduces a second dimension for one feature that is not present for this other (i.e., phoneme for Experiment 1 and voice identity plus gender for Experiment 2). Hence, couldn't it also be that the infants extracted the regularity with which one gender voice followed the other? For instance, in List B, in the words, one gender is always followed by the other (M-F or F-M), while in 2/3 of the part-words, the gender is repeated (F-F and M-M). Wouldn't you expect the same pattern of results if infants learned regularities based on gender rather than identity?*

We used three female and three male voices to maximise acoustic variability. The streams were synthesised using MBROLA, which provides a limited set of artificial voices. Indeed, there were not enough French voices of acceptable quality, so we also used two Italian voices (the phonemes used existed in both Italian and French).

Voices differ in timbre, and female voices tend to be higher pitched. However, it is sometimes difficult to categorise low-pitched female voices and high-pitched male voices. Given that gender may be an important factor in infants' speech perception (newborns, for instance, prefer female voices at birth), we conducted tests to assess whether this dimension could have influenced our results.

We report these analyses in SI and referred to them in the methods section (p.25, l.468 “We performed post-hoc tests to ensure that the results were not driven by a perception of two voices: female and male (see SI).”).

We first quantified the transitional probabilities matrices during the structured stream of Experiment 2, considering that there are only two types of voices: Female and Male.

For List A, all transition probabilities are equal to 0.5 (P(M|F), P(F|M), P(M|M), P(F|F)), resulting in flat TPs throughout the stream (see Author response image 1, top). Therefore, we

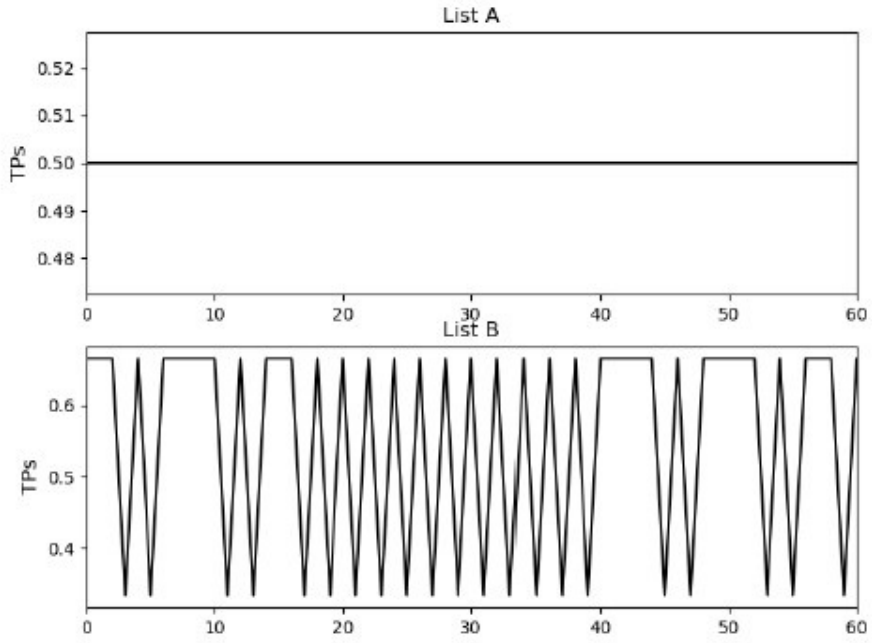
would not expect neural entrainment at the word rate (2 Hz), nor would we anticipate ERP differences between the presented duplets in the test phase.

For List B, P(M|F)=P(F|M)=0.66 while P(M|M)=P(F|F)=0.33. However, this does not produce a regular pattern of TP drops throughout the stream (see Author response image 1, bottom). As a result, strong neural entrainment at 2 Hz was unlikely, although some degree of entrainment might have occasionally occurred due to some drops occurring at a 2 Hz frequency. Regarding the test phase, all three Words and only one Part-word presented alternating patterns (TP=0.6). Therefore, the difference in the ERPs between Words and Part-words in List B might be attributed to gender alternation.

However, it seems unlikely that gender alternation alone explains the entire pattern of results, as the effect is inconsistent and appears in only one of the lists. To rule out this possibility, we analysed the effects in each list separately.

Author response image 1.

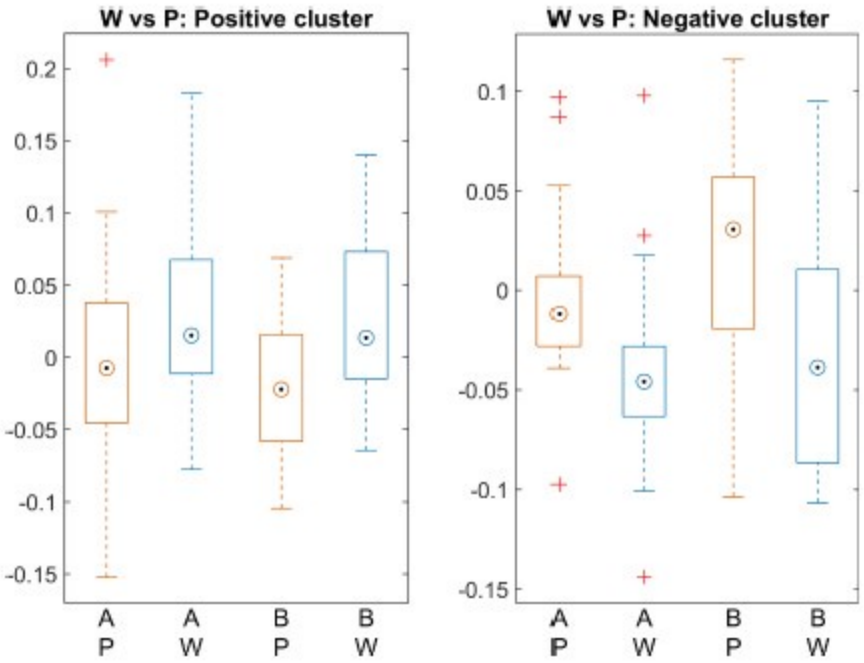
Transition probabilities (TPs) across the structured stream in Experiment 2, considering voices processed by gender (Female or Male). Top: List A. Bottom: List B.



We computed the mean activation within the time windows and electrodes of interest and compared the effects of word type and list using a two-way ANOVA. For the difference between Words and Part-words over the positive cluster, we observed a main effect of word type ($F(1,31) = 5.902$, $p = 0.021$), with no effects of list or interactions ($p > 0.1$). Over the negative cluster, we again observed a main effect of word type ($F(1,31) = 10.916$, $p = 0.0016$), with no effects of list or interactions ($p > 0.1$). See Author response image 2.

Author response image 2:

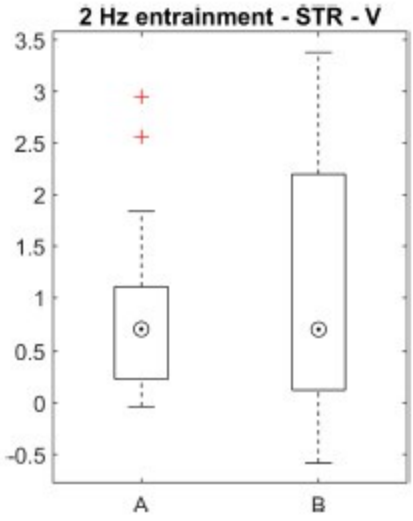
Difference in ERP voltage (Words – Part-words) for the two lists (A and B); W=Words; P=Part-Words,



We conducted a similar analysis for neural entrainment during the structured stream on voices. A comparison of entrainment at 2 Hz between participants who completed List A and List B showed no significant differences (t(30) = -0.27, p = 0.79). A test against zero for each list indicated significant entrainment in both cases (List A: t(17) = 4.44, p = 0.00036; List B: t(13) = 3.16, p = 0.0075). See Author response image 3.

Author response image 3.

Neural entrainment at 2Hz during the structured stream of Experiment 2 for Lists A and B.



> *Words entrainment over occipital electrodes: Do you have any idea why the duplet entrainment effect occurs over the electrodes it does, in particular over the occipital*

electrodes (which seems a bit unintuitive given that this is a purely auditory experiment with sleeping neonates).

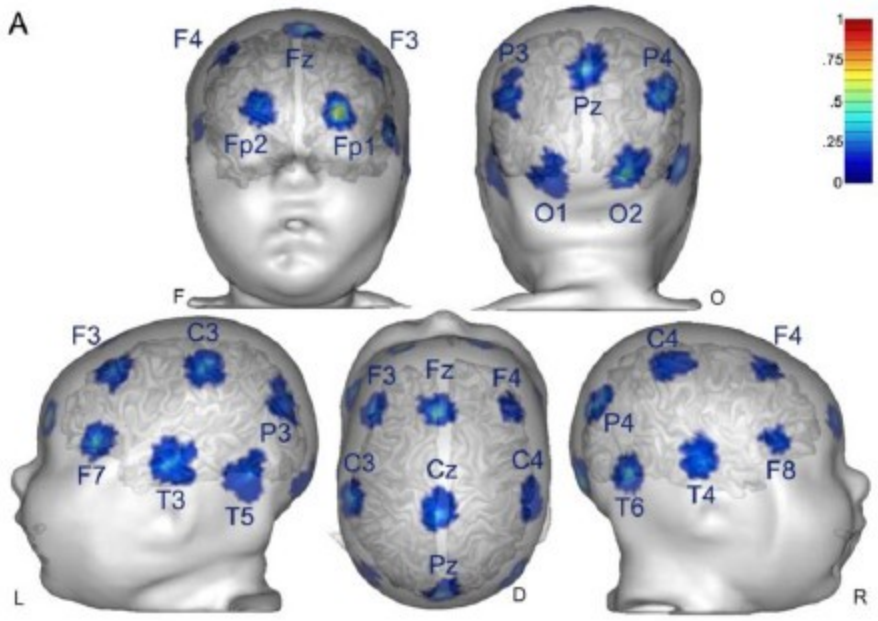
Neural entrainment might be considered as a succession of evoked response induced by the stream. After applying an average reference in high-density EEG recordings, the auditory ERP in neonates typically consists of a central positivity and a posterior negativity with a source located at the electrical zero in a single-dipole model (i.e. approximately in the superior temporal region (Dehaene-Lambertz & Dehaene, 1994). In adults, because of the average reference (i.e. the sum of voltages is equal to zero at each time point) and because the electrodes cannot capture the negative pole of the auditory response, the negativity is distributed around the head. In infants, however, the brain is higher within the skull, allowing for a more accurate recording of the negative pole of the auditory ERP (see Figure 4 for the location of electrodes in an infant head model).

Besides the posterior electrodes, we can see some entrainment on more anterior electrodes that probably corresponds to the positive pole of the auditory ERP.

We added a phrase in the discussion to explain why we can expect phase-locked activity in posterior electrodes (p.14, l.277: “Auditory ERPs, after reference-averaged, typically consist of a central positivity and posterior negativity”).

Author response image 4:

International 10–20 sensors' location on the skull of an infant template, with the underlying 3-D reconstruction of the grey-white matter interface and projection of each electrode to the cortex. Computed across 16 infants (from Kabdebon et al, Neuroimage, 2014). The O1, O2, T5, and T6 electrodes project lower than in adults.



Response to Reviewer 3:

(1) While it's true that voice is not essential for language (i.e., sign languages are implemented over gestures; the use of voices to produce non-linguistic sounds, like laughter), it is a feature of spoken languages. Thus I'm not sure if we can really consider

> *this study as a comparison between linguistic and non-linguistic dimensions. In turn, I'm not sure that these results show that statistical learning at birth operates on non-linguistic features, being voices a linguistic dimension at least in spoken languages. I'd like to hear the authors' opinions on this.*

On one hand, it has been shown that statistical learning (SL) operates across multiple modalities and domains in human adults and animals. On the other hand, SL is considered essential for infants to begin parsing speech. Therefore, we aimed to investigate whether SL capacities at birth are more effective on linguistic dimensions of speech, potentially as a way to promote language learning.

We agree with the reviewer that voices play an important role in communication (e.g., for identifying who is speaking); however, they do not contribute to language structure or meaning, and listeners are expected to normalize across voices to accurately perceive phonemes and words. Thus, voices are speech features but not linguistic features. Additionally, in natural speech, there are no abrupt voice changes within a word as in our experiment; instead, voice changes typically occur on a longer timescale and involve only a limited number of voices, such as in a dialogue. Therefore, computing regularities based on voice changes would not be useful in real-life language learning. We considered that contrasting syllables and voices was an elegant way to test SL beyond its linguistic dimension, as the experimental paradigm is identical in both experiments.

We have rephrased the introduction to make this point clearer. See p.5, l.88-92: “To test this, we have taken advantage of the fact that syllables convey two important pieces of information for humans: what is being said and who is speaking, i.e. linguistic content and speaker’s identity. While statistical learning…”.

> *Along the same line, in the Discussion section, the present results are interpreted within a theoretical framework showing statistical learning in auditory non-linguistic (string of tones, music) and visual domains as well as visual and other animal species. I'm not sure if that theoretical framework is the right fit for the present results.*
>
> *(2) I'm not sure whether the fact that we see parallel and independent tracking of statistics in the two dimensions of speech at birth indicates that newborns would be able to do so in all the other dimensions of the speech. If so, what other dimensions are the authors referring to?*

The reviewer is correct that demonstrating the universality of SL requires testing additional modalities and acoustic dimensions. However, we postulate that SL is grounded in a basic mechanism of long-term associative learning, as proposed in Benjamin et al. (2024), which relies on a slow decay in the representation of a given event. This simple mechanism, capable of operating on any representational output, accounts for many types of sequence learning reported in the literature (Benjamin et al., in preparation).

We have revised the discussion to clarify this theoretical framework.

In p.13, l.264: “This mechanism might be rooted in associative learning processes relying on the co- existence of event representations driven by slow activation decays (Benjamin et al., 2024). ”

In p., l. 364: “Altogether, our results show that statistical learning works similarly on different speech features in human neonates with no clear advantage for computing linguistically relevant regularities in speech. This supports the idea that statistical learning is a general learning mechanism, probably operating on common computational principles across neural networks (Benjamin et al., 2024)…”.

> *(3) Lines 341-345: Statistical learning is an evolutionary ancient learning mechanism but I do not think that the present results are showing it. This is a study on human neonates and adults, there are no other animal species involved therefore I do not see a connection with the evolutionary history of statistical learning. It would be much more interesting to make claims on the ontogeny (rather than philogeny) of statistical learning, and what regularities newborns are able to detect right after birth. I believe that this is one of the strengths of this work.*

We did not intend to make claims about the phylogeny of SL. Since SL appears to be a learning mechanism shared across species, we use it as a framework to suggest that SL may arise from general operational principles applicable to diverse neural networks. Thus, while it is highly useful for language acquisition, it is not specific to it.

We have removed the sentence “Statistical learning is an evolutionary ancient learning mechanism.”, and replaced it by (p.18, l.364) “Altogether, our results show that statistical learning works similarly on different speech features in human neonates with no clear advantage for computing linguistically relevant regularities in speech.” We now emphasise in the discussion that infants compute regularities on both features and propose that SL might be a universal learning mechanism sharing computational principles (Benjamin et al., 2024) (see point 2).

> *(4) The description of the stimuli in Lines 110-113 is a bit confusing. In Experiment 1, e.g., "pe" and "tu" are both uttered by the same voice, correct? ("random voice each time" is confusing). Whereas in Experiment 2, e.g., "pe" and "tu" are uttered by different voices, for example, "pe" by yellow voice and "tu" by red voice. If this is correct, then I recommend the authors to rephrase this section to make it more clear.*

To clarify, in Experiment 1, the voices were randomly assigned to each syllable, with the constraint that no voice was repeated consecutively. This means that syllables within the same word were spoken by different voices, and each syllable was heard with various voices throughout the stream. As a result, neonates had to retrieve the words based solely on syllabic patterns, without relying on consistent voice associations or specific voice relationships.

In Experiment 2, the design was orthogonal: while the syllables were presented in a random order, the voices followed a structured pattern. Similar to Experiment 1, each syllable (e.g., “pe” and “tu”) was spoken by different voices. The key difference is that in Experiment 2, the structured regularities were applied to the voices rather than the syllables. In other words, the “green” voice was always followed by the “red” voice for example but uttered different syllables.

We have revised the description of the stimuli and the legend of Figure 1 to clarify these important points.

See p.6, l. 113: “The structure consisted of the random concatenation of three duplets (i.e., two-syllable units) defined only by one of the two dimensions. For example, in Experiment 1, one duplet could be petu with each syllable uttered by a random voice each time they appear in the stream (e.g pe is produced by voice1 and tu by voice6 in one instance and in another instance pe is produced by voice3 and tu by

voice2). In contrast, in Experiment 2, one duplet could be the combination [voice1- voice6], each uttering randomly any of the syllables.”

p.20, l. 390 (Figure 1 legend): “For example, the two syllables of the word “petu” were produced by different voices, which randomly changed at each presentation of the word (e.g.

“yellow” voice and “green” voice for the first instance, “blue” and “purple” voice for the second instance, etc..). In Experiment 2, the statistical structure was based on voices (TPs alternated between 1 and 0.5), while the syllables changed randomly (uniform TPs of 0.2). For example, the “green” voice was always followed by the “red” voice, but they were randomly saying different syllables “boda” in the first instance, “tupe” in the second instance, etc... “

> *(5) Line 114: the sentence "they should compute a 36 x 36 TPs matrix relating each acoustic signal, with TPs alternating between 1/6 within words and 1/12 between words" is confusing as it seems like there are different acoustic signals. Can the authors clarify this point?*

Thank you for highlighting this point. To clarify, our suggestion is that neonates might not track regularities between phonemes and voices as separate features. Instead, they may treat each syllable-voice combination as a distinct item—for example, "pe" spoken by the "yellow" voice is one item, while "pe" spoken by the "red" voice is another. Under this scenario, there would be a total of 36 unique items (6 syllables × 6 voices), and infants would need to track regularities between these 36 combinations.

We have modified this sentence in the manuscript to make it clearer.

See p.7, l. 120: “If infants at birth compute regularities based on a neural representation of the syllable as a whole, i.e. comprising both phonetic and voice content, this would require computing a 36 × 36 TPs matrix relating each token.”

Reviewer #1 (Recommendations for the authors):

> *(1) The acronym TP should be spelled out, and a brief description of the fact that dips in TPs signal boundaries while high TPs signal a cohesive unit could be useful for non-specialist readers.*

We have added it at the beginning of the introduction (lines 52-60)

> *(2) p.5, l.76: "Here, we aimed to further characterise the characteristics of this mechanism...". I suggest this is rephrased as "to further characterise this mechanism".*

We have changed it as suggested by the reviewer (now p.5, l.81)

> *(3) p.9, l.172: "[...] this contribution is unlikely since the electrodes differ from the electrodes, showing enhanced word-rate activity at 2 Hz."*
>
> *It is unclear which electrodes differ from which electrodes. I figure that the authors mean that the electrodes showing stronger activity at 2 Hz differ from those showing it at 4 Hz, but the sentence could use rephrasing.*

This part has been rephrased (p.9, l.177-181)

> *(4) p.10, l.182: "[...] the entrainment during the first minute of the structure stream [...]".*
>
> *Structured stream.*

It has been corrected (p.10, l.190)

> *(5) p.12, l.234: "we compared STATISTICAL LEARNING"*
>
> *Why the use of capitals?*

This was an error and it was corrected (p.12, l.242).

> *(6) p.15, l.298: "[...] suggesting that such negativity might be related to semantic."*
>
> *The sentence feels incomplete. To semantics? To the processing of semantic information?*

The phrase has been corrected (p.15, l.314). Additionally, the discussion of the posterior negativity observed for duplets after familiarisation with a stream with regularities over phonemes has been rephrased (p.15, l.)

> *(7) Same page, l.301: "3-mo-olds" 3-month-olds.*

It has been corrected (now in p.16, l.333)

> *(8) Same page, l.307: "(see also (Bergelson and Aslin, 2017)" (see also Bergelson and Aslin, 2017).*

It has been corrected (now in p.17, l.340)

> *(9) Same page, l.310: "[...] would be considered as possible candidate" As possible candidates.*

This has been rephrased and corrected (now in p.17, l.343)

Reviewer #2 (Recommendations for the authors):

> *(1) Figure 2: The authors mention a "thick orange line", which I think should be a "thick black line".*

We are sorry for this. It has been corrected.

> *(2) Ln 166: Should be Figure 2C rather than 3C.*

It has been corrected (now in p.9, l.173)

> *(3) Figure 4 is not referenced in the manuscript.*

We referred to it now on p. 12, l.236

https://doi.org/10.7554/eLife.101802.2.sa0