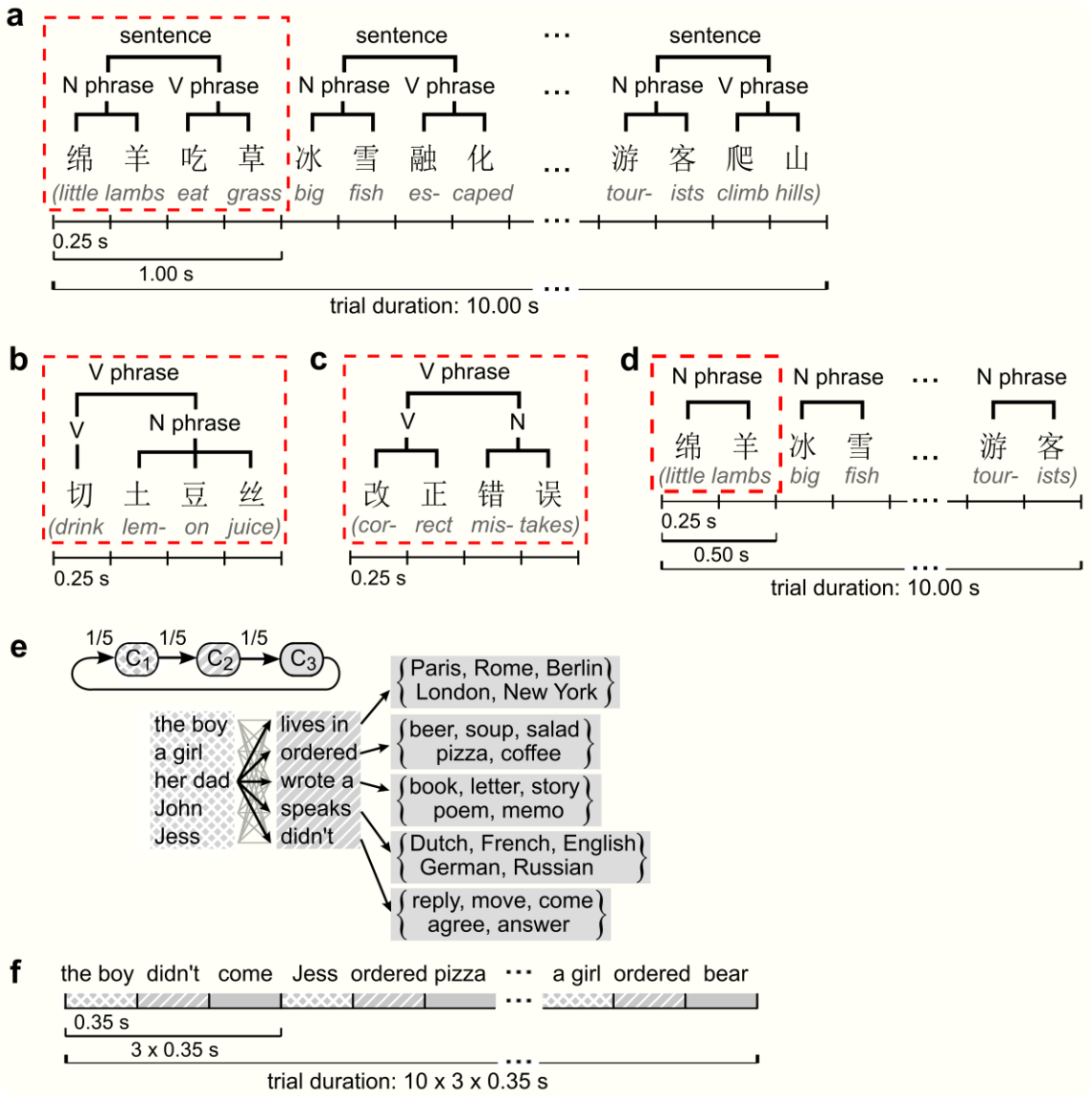
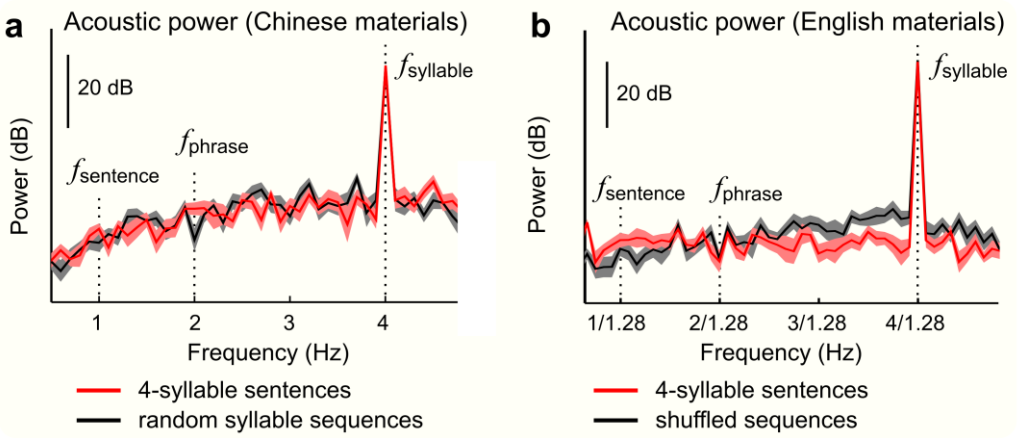


Supplementary Figure 1

Trial structure of Chinese (A-D) and English (EF) speech materials.

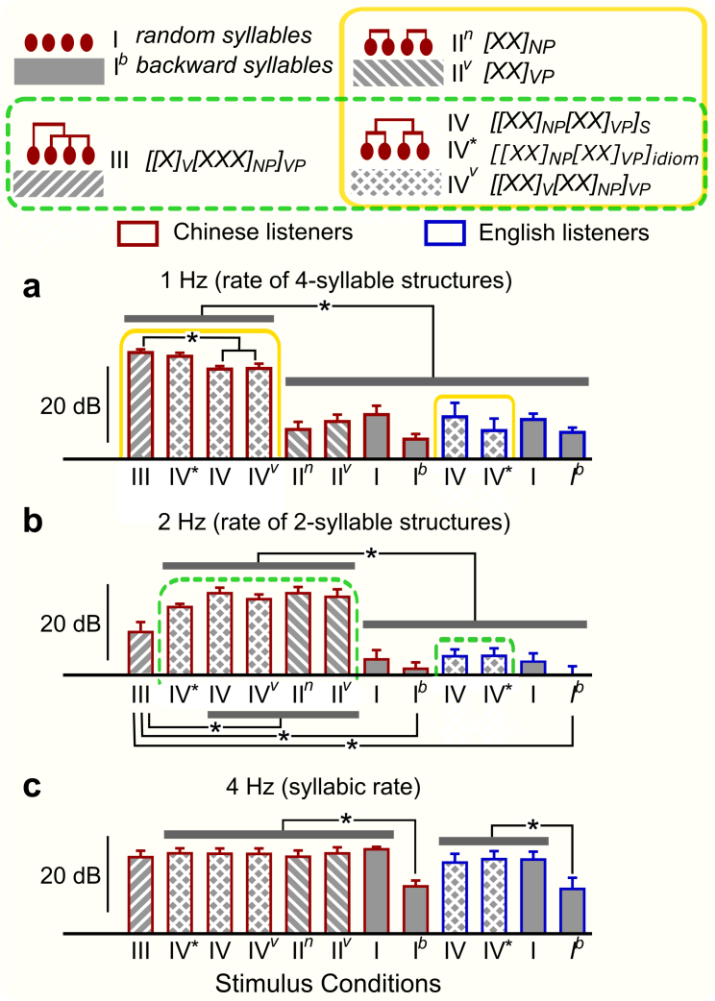
(A) For 4-syllable sentences, in each trial, 10 sentences are presented sequentially without any acoustic gap between them. English examples are given below the Chinese sentences/phrases to illustrate their syntactic structures (not direct translations). The same trial structure applies for 4-syllable verb phrases, except that each 4-syllable sentence (bounded by the dashed red box) is replaced by a 4-syllable type I verb phrase (B) or type II verb phrase (C). (D) For 2-syllable phrases, 20 phrases are presented sequentially in each trial. (E) Grammar for the constant predictability Markovian language. (F) The trial structure of Markovian language stimulus.



Supplementary Figure 2

The spectrum of the temporal envelope for the Chinese (A) and English (B) 4-syllable sentence stimuli.

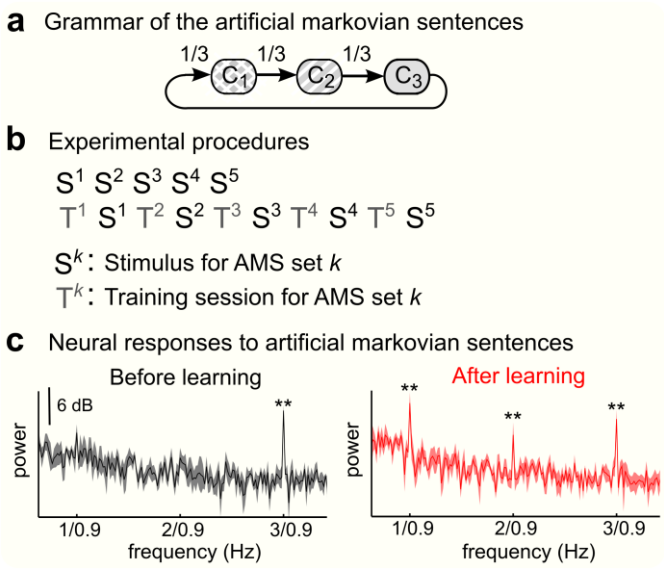
The power spectrum is averaged over all stimulus trials, and the SEM across trials is shown (shaded area). A spectral peak is seen at the syllabic rate but not at the phrasal or sentential rates, confirming that the sentential and phrasal structure is not conveyed by acoustic power cues. The stimulus envelope is the half-wave rectified sound waveform. The two conditions shown for each language are not significantly different from each other ($P > 0.15$, FDR corrected).



Supplementary Figure 3

Comparisons between the responses to stimuli of different linguistic structures.

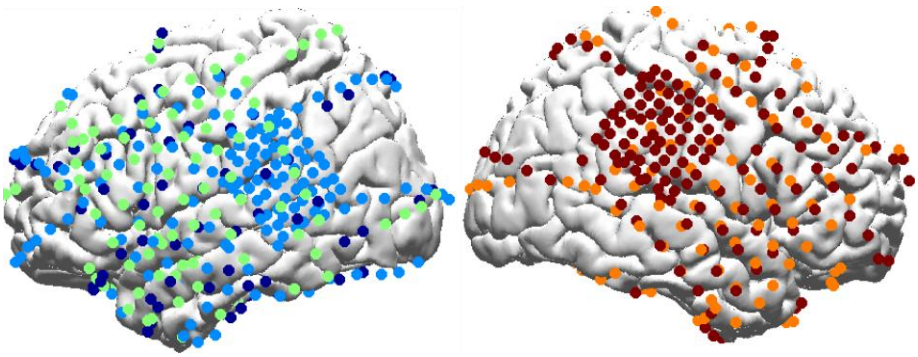
The tree diagrams at the top illustrate the four linguistic structures tested. All of them are constructed using an isochronous syllable sequence at 4 Hz. For Structure I, syllables or backward syllables are presented in a random order, not grouped into larger linguistic structures. For Structure II, every two syllables combine into a phrase, which activates a phrasal rhythm at 2Hz in addition to the 4-Hz syllabic rhythm. For Structure III, a 4-syllable verb phrase is constructed using a monosyllablic verb followed by a 3-syllable noun phrase. The 4-syllable verb phrase is frequency-tagged at 1 Hz but no linguistic structure is uniquely tagged at 2 Hz. For Structure IV, a 4-syllable structure evenly divides into two 2-syllable structures. The binary hierarchical embedding results in three levels of linguistic structures tagged at 1 Hz, 2 Hz, and 4 Hz, respectively. (A) For Chinese listeners (dark red bars), the 1-Hz response is significantly stronger for stimuli containing a 4-syllable constituent structure (yellow box). For English listeners who cannot parse the linguistic structure (blue bars), however, the response is not significantly different between conditions. All significant differences between conditions are shown and a thick gray bar indicates significant differences between two groups such that each condition in one group is significantly different from any condition in the other group ($P < 0.03$, t-test, FDR corrected). (B) The response at 2 Hz is stronger for stimuli containing 2-syllable phrasal structures (dashed green box) for Chinese listeners, but not so for English listeners. (C) A 4 Hz response, at the syllabic rate is seen in all tested conditions and both listener groups, but weaker for backward syllables than normal syllables.



Supplementary Figure 4

Dissociating neural encoding of sentential structures and transitional probability using Artifical Markovian Sentences (AMS).

(A) Grammar of the AMS. Each AMS consisted of 3 components, and each syllable was independently chosen from 3 candidate syllables with equal probability. In each trial, 33 sentences were played in a sequence without any gap in between them. (B) Procedures of the AMS experiment. The experiment has two sessions. In the first session (upper row), stimuli from each set of the AMS were played in separate blocks, before the listeners were instructed about the grammar of the AMS. In the second session, the 5 sets of AMS were learned in separate blocks. In the training phase of each block (labeled by T), the listeners listened to sentences from the AMS set and these sentences were separated by a 300 ms gap. After the training phase, the listeners listened to the same stimuli they heard in the first session. At the end of the block, the listeners had to report the grammar of the AMS set. (C) Neural response spectrum before (left) and after training (right). Before the listeners learn the grammar of the AMS, cortical activity only tracks the syllabic rhythm of speech. After learning, however, cortical activity concurrently follows the syllabic rhythm and the sentential rhythm. Since each trial (excluding the first sentence) is 53.1 seconds in duration, the frequency resolution of the spectrum is 0.019 Hz. Frequency bins showing power stronger than the mean power of a neighboring 1 Hz region (i.e., 0.5 Hz on each side) are shown by stars (N = 5, P < 0.001, paired t-test, FDR corrected).



Supplementary Figure 5

Coverage of the ECoG electrodes.

Color differentiates the 5 participants.