Cortical tracking of hierarchical linguistic structures in connected speech

Nai Ding[1,2], Lucia Melloni[3–5], Hang Zhang[1,6–8], Xing Tian[1,9,10] & David Poeppel[1,11]

The most critical attribute of human language is its unbounded combinatorial nature: smaller elements can be combined into larger structures on the basis of a grammatical system, resulting in a hierarchy of linguistic units, such as words, phrases and sentences. Mentally parsing and representing such structures, however, poses challenges for speech comprehension. In speech, hierarchical linguistic structures do not have boundaries that are clearly defined by acoustic cues and must therefore be internally and incrementally constructed during comprehension. We found that, during listening to connected speech, cortical activity of different timescales concurrently tracked the time course of abstract linguistic structures at different hierarchical levels, such as words, phrases and sentences. Notably, the neural tracking of hierarchical linguistic structures was dissociated from the encoding of acoustic cues and from the predictability of incoming words. Our results indicate that a hierarchy of neural processing timescales underlies grammar-based internal construction of hierarchical linguistic structure.

To understand connected speech, listeners must construct a hierarchy of linguistic structures of different sizes, including syllables, words, phrases and sentences[1–3]. It remains puzzling how the brain simultaneously handles the distinct timescales of the different linguistic structures, for example, from a few hundred milliseconds for syllables to a few seconds for sentences[4–14]. Previous studies have suggested that cortical activity is synchronized to acoustic features of speech, approximately at the syllabic rate, providing an initial timescale for speech processing[15–19]. But how the brain utilizes such syllabic-level phonological representations closely aligned with the physical input to build multiple levels of abstract linguistic structure, and represent these concurrently, is not known. We hypothesized that cortical dynamics emerge at all timescales required for the processing of different linguistic levels, including the timescales corresponding to larger linguistic structures such as phrases and sentences, and that the neural representation of each linguistic level corresponds to timescales matching the timescales of the respective linguistic level.

Although linguistic structure building can clearly benefit from prosodic[20,21] or statistical cues[22], it can also be achieved purely on the basis of the listeners' grammatical knowledge. To experimentally isolate the neural representation of the internally constructed hierarchical linguistic structure, we developed new speech materials in which the linguistic constituent structure was dissociated from prosodic or statistical cues. By manipulating the levels of linguistic abstraction, we found separable neural encoding of each different linguistic level.

RESULTS

Cortical tracking of phrasal and sentential structures

In the first set of experiments, we sought to determine the neural representation of hierarchical linguistic structure in the absence of prosodic cues. We constructed hierarchical linguistic structures using an isochronous, 4-Hz sequence of syllables that were independently synthesized (**Fig. 1a**,**b**, **Supplementary Fig. 1** and **Supplementary Table 1**). As a result of the acoustic independence between syllables (that is, no co-articulation), the linguistic constituent structure could only be extracted using lexical, syntactic and semantic knowledge, and not prosodic cues. The materials were first developed in Mandarin Chinese, in which syllables are relatively uniform in duration and are also the basic morphological unit (always morphemes and, in most cases, monosyllabic words). Cortical activity was recorded from native listeners of Mandarin Chinese using magnetoencephalography (MEG). Given that different linguistic levels, that is, the monosyllabic morphemes, phrases and sentences, were presented at unique and constant rates, the hypothesized neural tracking of hierarchical linguistic structure was tagged at distinct frequencies.

The MEG response was analyzed in the frequency domain and we extracted response power in every frequency bin using an optimal spatial filter (Online Methods). Consistent with our hypothesis, the response spectrum showed three peaks at the syllabic rate ($P = 1.4 \times 10^{-5}$, paired one-sided t test, false discovery rate (FDR) corrected), phrasal rate ($P = 1.6 \times 10^{-4}$, paired one-sided t test, FDR corrected) and sentential rate ($P = 9.6 \times 10^{-7}$, paired one-sided t test, FDR

[1]Department of Psychology, New York University, New York, New York, USA. [2]College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China. [3]Department of Neurology, New York University Langone Medical Center, New York, New York, USA. [4]Department of Neurophysiology, Max-Planck Institute for Brain Research, Frankfurt, Germany. [5]Department of Psychiatry, Columbia University, New York, New York, USA. [6]Department of Psychology and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China. [7]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China. [8]Peking-Tsinghua Center for Life Sciences, Beijing, China. [9]New York University Shanghai, Shanghai, China. [10]NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, China. [11]Neuroscience Department, Max-Planck Institute for Empirical Aesthetics, Frankfurt, Germany. Correspondence should be addressed to N.D. (ding_nai@zju.edu.cn) or D.P. (david.poeppel@nyu.edu).

Received 12 August; accepted 3 November; published online 7 December 2015; doi:10.1038/nn.4186

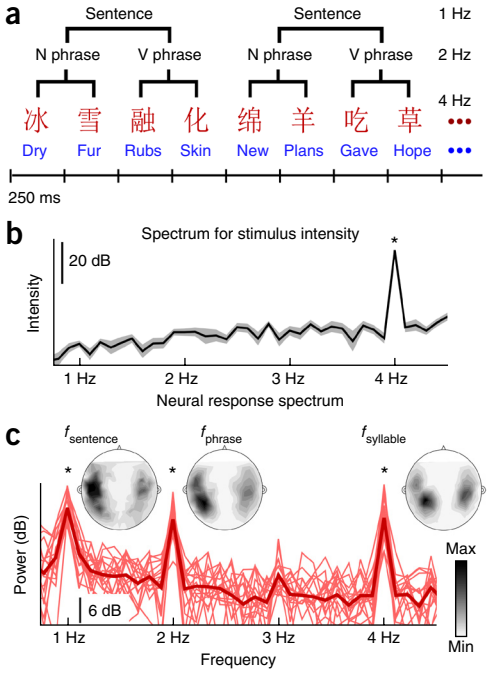
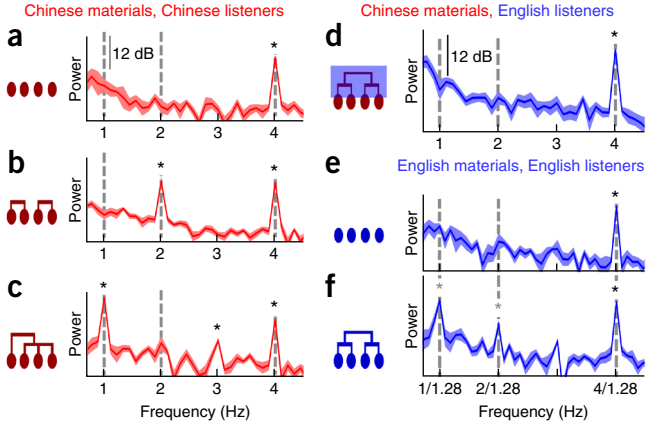


Figure 1 Neural tracking of hierarchical linguistic structures. (**a**) Sequences of Chinese or English monosyllabic words were presented isochronously, forming phrases and sentences. (**b**) Spectrum of stimulus intensity fluctuation revealed syllabic rhythm, but no phrasal or sentential modulation. The shaded area covers 2 s.e.m. across stimuli. (**c**) MEG-derived cortical response spectrum for Chinese listeners and materials (dark red curve, grand average; light red curves, individual listeners; $N = 16$, 0.11-Hz frequency resolution). Neural tracking of syllabic, phrasal and sentential rhythms was reflected by spectral peaks at corresponding frequencies. Frequency bins with significantly stronger power than neighbors (0.5 Hz range) are marked (*$P < 0.001$, paired one-sided t test, FDR corrected). The topographical maps of response power across sensors are shown for the peak frequencies.

corrected) and the response was highly consistent across listeners (**Fig. 1c**). Given that the phrasal- and sentential-rate rhythms were not conveyed by acoustic fluctuations at the corresponding frequencies (**Fig. 1b**), cortical responses at the phrasal and sentential rates must be a consequence of internal online structure building processes. Cortical activity at all the three peak frequencies was seen bilaterally (**Fig. 1c**). The response power averaged over sensors in each hemisphere was significantly stronger in the left hemisphere at the sentential rate ($P = 0.014$, paired two-sided t test), but not at the phrasal ($P = 0.20$, paired two-sided t test) or syllabic rates ($P = 0.40$, paired two-sided t test).

Dependence on syntactic structures

Are the responses at the phrasal and sentential rates indeed separate neural indices of processing at distinct linguistic levels or are they merely sub-harmonics of the syllabic rate response, generated by intrinsic cortical dynamical properties? We address this question by manipulating different levels of linguistic structure in the input. When the stimulus is a sequence of random syllables that preserves the acoustic properties of Chinese sentences (**Fig. 1** and **Supplementary Fig. 2**), but eliminates the phrasal/sentential structure, only syllabic (acoustic) level tracking occurs ($P = 1.1 \times 10^{-4}$ at 4 Hz, paired one-sided t test, FDR corrected; **Fig. 2a**). Furthermore, this manipulation preserves the position of each syllable in a sentence (Online Methods) and therefore further demonstrates that the phrasal- and sentential-rate responses are not a result of possible acoustic differences between the syllables in a sentence. When two adjacent syllables and morphemes combine into verb phrases, but there is no four-element sentential structure, phrasal-level tracking emerges at half of the syllabic rate ($P = 8.6 \times 10^{-4}$ at 2 Hz and $P = 2.7 \times 10^{-4}$ at 4 Hz, paired one-sided t test, FDR corrected; **Fig. 2b**). Similar responses are observed for noun phrases (**Supplementary Fig. 3**).



To test whether the phrase-level responses segregate from the sentence level, we constructed longer verb phrases that were unevenly divided into a monosyllabic verb followed by a three-syllable noun phrase (**Fig. 2c**). We expect that the neural responses to the long verb phrase to be tagged at 1 Hz, whereas the neural responses to the monosyllabic verb and the three-syllable noun phrase will present as harmonics of 1 Hz. Consistent with our hypothesis, cortical dynamics emerged at one-fourth of the syllabic rate, whereas the response at half of the syllabic rate is no longer detectable ($P = 1.9 \times 10^{-4}$, 1.7×10^{-4} and 9.3×10^{-4} at 1, 3 and 4 Hz, respectively, paired one-sided t test, FDR corrected).

Dependence on language comprehension

When listening to Chinese sentences (**Fig. 1a**), listeners who did not understand Chinese only showed responses to the syllabic (acoustic) rhythm ($P = 3.0 \times 10^{-5}$ at 4 Hz, paired one-sided t test, FDR corrected; **Fig. 2d**), further supporting the argument that cortical responses to larger, abstract linguistic structures is a direct consequence of language comprehension.

If aligning cortical dynamics to the time course of linguistic constituent structure is a general mechanism required for comprehension, it must apply across languages. Indeed, when native English speakers were tested with English materials (**Fig. 1a**), their cortical activity also followed the time course of larger linguistic structures, that is, phrases and sentences ($P = 4.1 \times 10^{-5}$, syllabic rate; **Fig. 2e**; $P = 3.9 \times 10^{-3}$,

Figure 2 Tracking of different linguistic structures. Each panel shows syntactic structure repeating in the stimulus (left) and the cortical response spectrum (right; shaded area indicates 2 s.e.m. over listeners, $N = 8$). (**a**) Chinese listeners, Chinese materials: syllables were syntactically independent and cortical activity encoded only acoustic and syllabic rhythm. (**b**,**c**) Additional tracking emerged with larger linguistic structures. Spectral peaks marked by a star (black, $P < 0.001$; gray, $P < 0.005$; paired one-sided t test, FDR corrected). (**d**) English listeners, Chinese materials from **Figure 1**: acoustic tracking only, as there was no parsable structure. (**e**,**f**) English listeners, English materials: syllabic rate (4/1.28 Hz) and sentential and phrasal rate responses to parsable structure in stimulus.

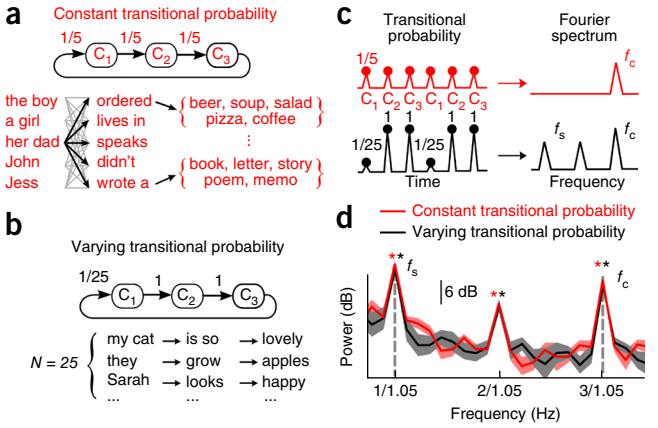


Figure 3 Dissociating sentential structures and transitional probability. (**a**,**b**) Grammar of an artificial Markovian stimulus set with constant (**a**) or variable (**b**) transitional probability. Each sentence consists of three acoustic chunks, each containing 1–2 English words. The listeners memorized the grammar before experiments. (**c**) Schematic time course and spectrum of the transitional probability. (**d**) Neural response spectrum (shaded area covers 2 s.e.m. over listeners, $N = 8$). Significant neural responses to sentences were seen for both languages. Spectral peaks are shown by an asterisk ($P < 0.001$, paired one-sided t test, FDR corrected, same color code as the spectrum). Responses were not significantly different between the two languages in any frequency bin (paired two-sided t test, $P > 0.09$, uncorrected).

4.3×10^{-3} and 6.8×10^{-6} at the sentential, phrasal and syllabic rates, respectively; **Fig. 2f**; paired one-sided t test, FDR corrected).

Neural tracking of linguistic structures rather than probability cues

We found that concurrent neural tracking of multiple levels of linguistic structure was not confounded with the encoding of acoustic cues (**Figs. 1** and **2**). However, is this simply explained by the neural tracking of the predictability of smaller units? As a larger linguistic structure, such as a sentence, unfolds in time, its component units become more predictable. Thus, cortical networks solely tracking transitional probabilities across smaller units could show temporal dynamics matching the timescale of larger structures. To test this alternative hypothesis, we crafted a constant transitional probability Markovian Sentence Set (MSS) in which the transitional probability of lower level units was dissociated from the higher level structures (**Fig. 3a** and **Supplementary Fig. 1e**,**f**). The constant transitional probability MSS is contrasted with a varying transitional probability MSS, in which the transitional probability is low across sentential boundaries and high in a sentence (**Fig. 3b**,**c**). If cortical activity only encodes the transitional probability between lower level units (for example, acoustic chunks in the MSS) independent of the underlying syntactic structure, it can show tracking of the sentential structure for the varying probability MSS, but not for the constant probability MSS. In contrast with this prediction, indistinguishable neural responses to sentences were observed for both MSS (**Fig. 3d**), demonstrating that neural tracking of sentences is not confounded by transitional probability. Specifically, for the constant transitional probability MSS, the response was statistically significant at the sentential rate, twice the sentential rate and the syllable rate ($P = 1.8 \times 10^{-4}$, 2.3×10^{-4} and 2.7×10^{-6}, respectively). For the varying transitional probability MSS, the response was statistically significant at the sentential rate, twice the sentential rate and the syllable rate ($P = 7.1 \times 10^{-4}$, 7.1×10^{-4} and 4.8×10^{-6}, respectively).

Given that the MSS involved real English sentences, listeners had prior knowledge of the transitional probabilities between acoustic chunks. To control for the effect of such prior knowledge, we created a set of Artificial Markovian Sentences (AMS). In the AMS, the transitional probability between syllables was the same in and across sentences (**Supplementary Fig. 4a**). The AMS was composed of Chinese syllables, but no meaningful Chinese expressions were embedded in the AMS sequences. As the AMS was not based on the grammar of Chinese, the listeners had to learn the AMS grammar to segment sentences. By comparing the neural responses to the AMS sequences before and after the grammar was learned, we were able to separate the effect of prior knowledge of transitional probability and the effect of grammar learning. Here, the grammar of the AMS indicates the set of rules that governs the sequencing of the AMS, that is, the rule of which syllables can follow which syllables.

The neural responses to the AMS before and after grammar learning were analyzed separately (**Supplementary Fig. 4**). Before learning, when the listeners were instructed that the stimulus was just a sequence of random syllables, the response showed a statistically significant peak at the syllabic rate ($P = 0.0003$, bootstrap), but not at the sentential rate. After the AMS grammar was learned, however, a significant response peak emerged at the sentential rate ($P = 0.0001$, bootstrap). A response peak was also observed at twice the sentential rate, possibly reflecting the second harmonic of the sentential response. This result further confirms that neural tracking of sentences is not confounded by neural tracking of transitional probability.

Neural tracking of sentences varying in duration and structure

These results are based on sequences of sentences that have uniform duration and syntactic structure. We next addressed whether cortical

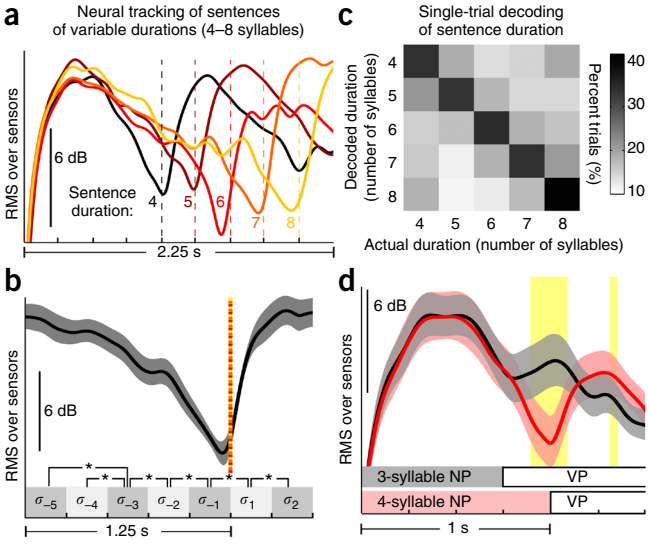


Figure 4 Neural tracking of sentences of varying structures. (**a**) Neural activity tracked the sentence duration, even when the sentence boundaries (dotted lines) were not separated by acoustic gaps. (**b**) Averaged response near a sentential boundary (dotted line). The power continuously changed throughout the duration of a sentence. Shaded area covers 2 s.e.m. over listeners ($N = 8$). Significance power differences between time bins (shaded squares) are marked by asterisks ($P = 0.01$, one-sided t test, FDR corrected). (**c**) Confusion matrix for neural decoding of the sentence duration. (**d**) Neural activity tracks noun phrase duration (shown in the bottom). Yellow areas show significant differences between curves ($P = 0.005$, bootstrap, FDR corrected).

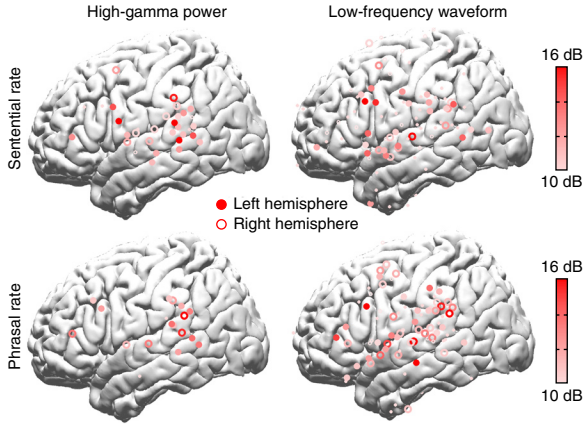


Figure 5 Localizing cortical sources of the sentential and phrasal rate responses using ECoG ($N = 5$). Left, power envelope of high-gamma activity. Right, waveform of low-frequency activity. Electrodes in the right hemisphere were projected to the left hemisphere, and right hemisphere (left hemisphere) electrodes are shown by hollow (filled) circles. The figure only displays electrodes that showed statistically significant neural responses to sentences in **Figure 2e** and no significant response to the acoustic control shown in **Figure 2f**. Significance was determined by bootstrap (FDR corrected) and the significance level is 0.05. The response strength, that is, the response at the target frequency relative to the mean response averaged over a 1-Hz wide neighboring region, is color coded. Electrodes with response strength less than 10 dB are shown by smaller symbols. The sentential and phrasal rate responses were seen in bilateral pSTG, TPJ and left IFG.

tracking of larger linguistic structures generalizes to sentences that are variable in duration (4–8 syllables) and syntactic structures. These sentences were again built on isochronous Chinese syllables, intermixed and sequentially presented without any acoustic gap at the sentence boundaries. Examples translated into English include "Don't be nervous," "The book is hard to read," and "Over the street is a museum."

As these sentences have irregular durations that are not tagged by frequency, the MEG responses were analyzed in the time domain by averaging sentences of the same duration. To focus on sentential level processing, we low-pass filtered the response at 3.5 Hz. The MEG response (root mean square, r.m.s., over all sensors) rapidly increased after a sentence boundary and continuously changed throughout the duration of a sentence (**Fig. 4a**). To illustrate the detailed temporal dynamics, we averaged the r.m.s. response over all sentences containing six or more syllables after aligning them to the sentence offset (**Fig. 4b**). During the last four syllables of a sentence, the r.m.s. response continuously and significantly decreased for every syllable, indicating that the neural response continuously changes during the course of a sentence rather than being a transient response only occurring at the sentence boundary.

A single-trial decoding analysis was performed to independently confirm that cortical activity tracks the duration of sentences (**Fig. 4c**). The decoder applied template matching for the response time course (leave-one-out cross-validation, Online Methods) and correctly determined the duration of 34.9 ± 0.6% sentences (mean ± s.e.m. over subjects, significantly above chance, $P = 1.3 \times 10^{-7}$, one-sided t test).

After demonstrating cortical tracking of sentences, we further tested whether cortical activity also tracks the phrasal structure inside of a sentence. We constructed sentences that consist of a noun phrase followed by a verb phrase and manipulated the duration of the noun phrase (three syllable or four syllable). The cortical responses closely follow the duration of the noun phrase: the r.m.s. response gradually decreased in the noun phrase, then showed a transient increase after the onset of the verb phrase (**Fig. 4d**).

Neural source localization using electrocorticography (ECoG)

We found that large-scale neural activity measured by MEG concurrently follows the hierarchical linguistic structure of speech, but which neural networks generate such activity? To address this question, we recorded the ECoG responses to English sentences (**Fig. 2e**) and an acoustic control (**Fig. 2f**). ECoG signals are mesoscopic neurophysiological signals recorded by intracranial electrodes implanted in

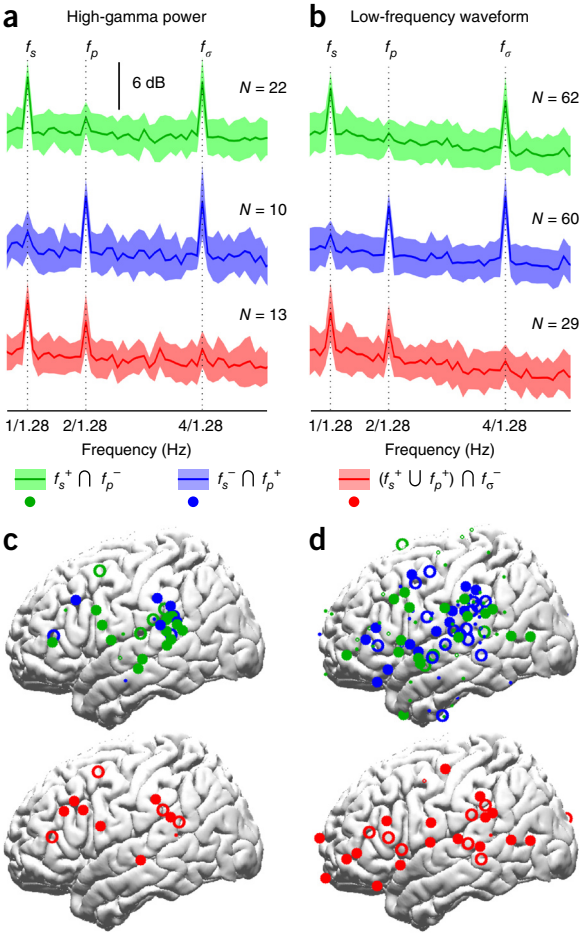


Figure 6 Spatial dissociation between sentential-rate, phrasal-rate and syllabic-rate responses ($N = 5$). (**a**) The power spectrum of the power envelope of high-gamma activity. (**b**) The power spectrum of low-frequency ECoG waveform. The top panels (green curves) show the response averaged over all electrodes that show a significant sentential-rate response but not a significant phrasal-rate response. Significance was determined by bootstrap (FDR corrected) and the significance level is 0.05. The shaded area is 1 s.d. over electrodes on each side. The blue curves show the response averaged over all electrodes that showed a significant phrasal-rate response, but not a significant sentential-rate response. The red curves show a significant sentential-rate or a significant phrasal-rate response, but not a significant syllabic response. (**c**,**d**) The topographic distribution of the three groups of electrodes analyzed in **a** and **b**. As in **Figure 5**, electrodes showing a response greater than 10 dB are shown by larger symbols than electrodes showing a response weaker than 10 dB.

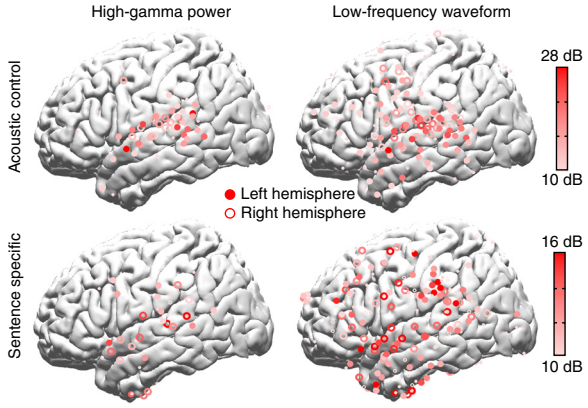


Figure 7 Syllabic-rate ECoG responses to English sentences and the acoustic control ($N = 5$). Top, electrodes showing statistically significant syllabic-rate ECoG responses to the acoustic control, that is, shuffled sequences, which had the same acoustic and syllabic rhythm as the English sentences, but contained no hierarchical linguistic structures (**Fig. 2f**). Significance was determined by bootstrap (FDR corrected) and the significance level is 0.05. The responses were most strongly seen in bilateral STG for both high-gamma and low-frequency activity and in bilateral pre-motor areas for low-frequency activity. Bottom, syllabic-rate ECoG responses to English sentences. The electrodes displayed are those that showed statistically significant neural responses to sentences and no significant response to the acoustic control. The syllabic rate responses specific to sentences were strong along bilateral STG for high-gamma activity and were widely distributed in the frontal and temporal lobes for low-frequency activity.

epilepsy patients for clinical evaluation (see **Supplementary Fig. 5** for the electrode coverage), and they possess better spatial resolution than MEG. We first analyzed the power of the ECoG signal in the high gamma band (70–200 Hz), as it highly correlates with multiunit firing[23]. The electrodes exhibiting significant sentential, phrasal and syllabic rate fluctuations in high gamma power are shown separately (**Fig. 5**). The sentential rate response clustered over the posterior and middle superior temporal gyrus (pSTG), bilaterally, with a second cluster over the left inferior frontal gyrus (IFG). Phrasal rate responses were also observed over the pSTG bilaterally. Notably, although the sentential and phrasal rate responses were observed in similar cortical areas, electrodes showing phrasal rate responses only partially overlapped with electrodes showing sentential rate responses in the pSTG (**Fig. 6**). For electrodes showing a significant response at either the sentential rate or the phrasal rate, the strength of the sentential rate response was negatively correlated with the strength of the phrasal rate response ($R = -0.32$, $P = 0.004$, bootstrap). This phenomenon demonstrates spatially dissociable neural tracking of the sentential and phrasal structures.

Furthermore, some electrodes with a significant sentential or phrasal rate response showed no significant syllabic rate response ($P < 0.05$, FDR corrected, **Fig. 6**). In other words, there are cortical circuits specifically encoding larger, abstract linguistic structures without responding to syllabic-level acoustic features of speech. In addition, although the syllabic responses were not significantly different ($P > 0.05$, FDR corrected) for English sentences and the acoustic control in the MEG results, they were dissociable spatially in the ECoG results (**Fig. 7**). Electrodes showing significant syllabic responses ($P < 0.05$, FDR corrected) to sentences, but not the acoustic control, were seen in bilateral pSTG, bilateral anterior STG (aSTG), and left IFG.

We then analyzed neural tracking of the sentential, phrasal and syllabic rhythms in the low-frequency ECoG waveform (**Fig. 5**), which is a close neural correlate of MEG activity. Fourier analysis was directly applied to the ECoG waveform and the Fourier coefficients at 1, 2 and 4 Hz are extracted. Low-frequency ECoG activity is usually viewed as the dendritic input to a cortical area[24]. The low-frequency responses are more distributed than high-gamma activity, possibly reflecting the fact that the neural representations of different levels of linguistic structures serve as inputs to broad cortical areas. Sentential and phrasal rate responses are strong in STG, IFG and temporoparietal junction (TPJ). Compared with the acoustic control, the syllabic-rate response to sentences was stronger in broad cortical areas, including the temporal and frontal lobes (**Fig. 7**). Similar to the high-gamma activity, the low-frequency responses to the sentential and phrasal structures were not reflected in the same set of electrodes (**Fig. 6**). For electrodes showing a significant response at either the sentential rate or the phrasal rate, the strength of the sentential rate response was also negatively correlated with the strength of the phrasal rate response ($R = -0.21$, significantly greater than 0, $P = 0.023$, bootstrap).

DISCUSSION

Our data show that the multiple timescales that are required for the processing of linguistic structures of different sizes emerge in cortical networks during speech comprehension. The neural sources for sentential, phrasal and syllabic rate responses are highly distributed and include cortical areas that have been found to be critical for prosodic (for example, right STG), syntactic and semantic (for example, left pSTG and left IFG) processing[9,25–28]. Neural integration on different timescales is likely to underlie the transformation from shorter lived neural representations of smaller linguistic units to longer lasting neural representations of larger linguistic structures[11–14]. These results underscore the undeniable existence of hierarchical structure building operations in language comprehension[1,2] and can be applied to objectively assess language processing in children and difficult-to-test populations, as well as animal preparations to allow for cross-species comparisons.

Relation to language comprehension

Concurrent neural tracking of hierarchical linguistic structures provides a plausible functional mechanism for temporally integrating smaller linguistic units into larger structures. In this form of concurrent neural tracking, the neural representation of smaller linguistic units is embedded at different phases of the neural activity tracking a higher level structure. Thus, it provides a possible mechanism to transform the hierarchical embedding of linguistic structures into hierarchical embedding of neural dynamics, which may facilitate information integration in time[10,11]. Low-frequency neural tracking of linguistic structures may further modulate higher frequency neural oscillations[29–31], which have been proposed to provide additional roles for structure building[7]. In addition, multiple resources and computations are needed for syntactic analysis, for example, access to combinatorial syntactic subroutines, and such operations may correspond to neural computations on distinct frequency scales, which are coordinated by the low-frequency neural tracking of linguistic constituent structures. Furthermore, low-frequency neural activity and oscillations have been hypothesized as critical mechanisms to generate predictions about future events[32]. For language processing, it is likely that concurrent neural tracking of hierarchical linguistic structures provides mechanisms to generate predictions on multiple linguistic levels and allow interactions across linguistic levels[33].

Neural entrainment to speech

Recent work has shown that cortex tracks the slow acoustic fluctuations of speech below 10 Hz (refs. 15–18,34,35), and this phenomenon is commonly described as 'cortical entrainment' to the syllabic rhythm of speech. It has been controversial whether such syllabic-level cortical entrainment is related to low-level auditory encoding or language processing[6]. Our findings demonstrate that processing goes well beyond stimulus-bound analysis: cortical activity is entrained to larger linguistic structures that are, by necessity, internally constructed, based on syntax. The emergence of slow cortical dynamics provides timescales suitable for the analysis of larger chunk sizes[13,14].

A long-lasting controversy concerns how the neural responses to sensory stimuli are related to intrinsic, ongoing neural oscillations. This question is heavily debated for the neural response entrained to the syllabic rhythm of speech[36] and can also be asked for neural activity entrained to the time courses of larger linguistic structures. Our experiment was not designed to answer this question; however, we clearly found that cortical speech processing networks have the capacity to generate activity on very long timescales corresponding to larger linguistic structures, such as phrases and sentences. In other words, the timescales of larger linguistic structures fall in the timescales, or temporal receptive windows[12,13], that the relevant cortical networks are sensitive to. Whether the capacity of generating low-frequency activity during speech processing is the same as the mechanisms generating low-frequency spontaneous neural oscillations will need to be addressed in the future.

Nature of the linguistic representations

Language processing is not monolithic and involves partially segregated cortical networks for the processing of, for example, phonological, syntactic and semantic information[9]. The phonological, syntactic and semantic representations are all hierarchically organized[37] and may rely on the same core structure building operations[38]. In natural speech, linguistic structure building can be facilitated by prosodic[39] and statistical cues[22], and some underlying neural signatures have been demonstrated[6,8,20]. Such cues, however, are not always available, and even when they are available, they are generally not sufficient. Thus, robust structure building relies on a listeners' tacit syntactic knowledge, and our findings provide unique insights into the neural representation of abstract linguistic structures that are internally constructed on the basis of syntax alone. Although the construction of abstract structures is driven by syntactic analysis, when such structures are built, different aspects of the structure, including semantic information, can be integrated in the neural representation. Indeed, the wide distribution of cortical tracking of hierarchical linguistic structures suggests that it is a general neurophysiological mechanism for combinatorial operations involved in hierarchical linguistic structure building in multiple linguistic processing networks (for example, phonological, syntactic and semantic). Furthermore, coherent synchronization to the correlated linguistic structures in different representational networks, for example, syntactic, semantic and phonological, provides a way to integrate multi-dimensional linguistic representations into a coherent language percept[38,40], just as temporal synchronization between cortical networks provides a possible solution to the binding problem in sensory processing[41].

Relation to event-related responses

Although many previous neurophysiological studies on structure building have focused on syntactic and semantic violations[42–44], fewer have addressed normal structure building; on the lexical-semantic level, the N400/N400m has been identified as a marker of the semantic predictability of words[43,45] and its amplitude continuously reduces in a sentence[46,47]. For syntactic processing, when two words combine into a short phrase, increased activity is seen in the temporal and frontal lobes[4]. Our results build on and extends these findings by demonstrating structure building at different levels of the linguistic hierarchy, during online comprehension of connected speech materials in which the structural boundaries are neither physically cued nor confounded by the semantic predictability of the individual words (**Fig. 3**). Note that, although the two Markovian languages (compared in **Fig. 3**) differed in their transitional probability between acoustic chunks, they both had fully predictable syntactic structures. The equivalence in syntactic predictability is likely to result in the very similar responses between the two conditions.

Lastly, the emergence of slow neural dynamics tracking superordinate stimulus structures is reminiscent of what has been observed during decision making[48], action planning[49] and music perception[50], suggesting a plausible common neural computational framework to integrate information over distinct timescales[12]. These findings invite MEG and EEG studies to extend from the classic event-related designs to investigating continuous neural encoding of internally constructed perceptual organization of an information stream.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Walker for MEG technical support, T. Thesen, W. Doyle and O. Devinsky for their instrumental help in collecting ECoG data, and G. Buzsaki, G. Cogan, S. Dehaene, A.-L. Giraud, G. Hickok, N. Hornstein, E. Lau, A. Marantz, N. Mesgarani, M. Peña, B. Pesaran, L. Pylkkänen, C. Schroeder, J. Simon and W. Singer for their comments on previous versions of the manuscript. This work supported by US National Institutes of Health grant 2R01DC05660 (D.P.) and Major Projects Program of the Shanghai Municipal Science and Technology Commission (STCSM) 15JC1400104 (X.T.) and National Natural Science Foundation of China 31500914 (X.T.).

AUTHOR CONTRIBUTIONS

N.D., L.M. and D.P. conceived and designed the experiments. N.D., H.Z. and X.T. performed the MEG experiments. L.M. performed the ECoG experiment. N.D., L.M. and D.P. wrote the paper. All of the authors discussed the results and edited the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Berwick, R.C., Friederici, A.D., Chomsky, N. & Bolhuis, J.J. Evolution, brain, and the nature of language. *Trends Cogn. Sci.* **17**, 89–98 (2013).
2. Chomsky, N. *Syntactic Structures* (Mouton de Gruyter, 1957).
3. Phillips, C. Linear order and constituency. *Linguist. Inq.* **34**, 37–90 (2003).
4. Bemis, D.K. & Pylkkänen, L. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cereb. Cortex* **23**, 1859–1873 (2013).
5. Giraud, A.-L. & Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
6. Sanders, L.D., Newport, E.L. & Neville, H.J. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. *Nat. Neurosci.* **5**, 700–703 (2002).
7. Bastiaansen, M., Magyari, L. & Hagoort, P. Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *J. Cogn. Neurosci.* **22**, 1333–1347 (2010).
8. Buiatti, M., Peña, M. & Dehaene-Lambertz, G. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage* **44**, 509–519 (2009).

9. Pallier, C., Devauchelle, A.-D. & Dehaene, S. Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci. USA* **108**, 2522–2527 (2011).
10. Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S. & Puce, A. Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* **12**, 106–113 (2008).
11. Buzsáki, G. Neural syntax: cell assemblies, synapsembles and readers. *Neuron* **68**, 362–385 (2010).
12. Bernacchia, A., Seo, H., Lee, D. & Wang, X.-J. A reservoir of time constants for memory traces in cortical neurons. *Nat. Neurosci.* **14**, 366–372 (2011).
13. Lerner, Y., Honey, C.J., Silbert, L.J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
14. Kiebel, S.J., Daunizeau, J. & Friston, K.J. A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* **4**, e1000209 (2008).
15. Luo, H. & Poeppel, D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* **54**, 1001–1010 (2007).
16. Ding, N. & Simon, J.Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. USA* **109**, 11854–11859 (2012).
17. Zion Golumbic, E.M. *et al.* Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **77**, 980–991 (2013).
18. Peelle, J.E., Gross, J. & Davis, M.H. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* **23**, 1378–1387 (2013).
19. Pasley, B.N. *et al.* Reconstructing speech from human auditory cortex. *PLoS Biol.* **10**, e1001251 (2012).
20. Steinhauer, K., Alter, K. & Friederici, A.D. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nat. Neurosci.* **2**, 191–196 (1999).
21. Peña, M., Bonatti, L.L., Nespor, M. & Mehler, J. Signal-driven computations in speech processing. *Science* **298**, 604–607 (2002).
22. Saffran, J.R., Aslin, R.N. & Newport, E.L. Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
23. Ray, S. & Maunsell, J.H. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9**, e1000610 (2011).
24. Einevoll, G.T., Kayser, C., Logothetis, N.K. & Panzeri, S. Modeling and analysis of local field potentials for studying the function of cortical circuits. *Nat. Rev. Neurosci.* **14**, 770–785 (2013).
25. Hagoort, P. & Indefrey, P. The neurobiology of language beyond single words. *Annu. Rev. Neurosci.* **37**, 347–362 (2014).
26. Grodzinsky, Y. & Friederici, A.D. Neuroimaging of syntax and syntactic processing. *Curr. Opin. Neurobiol.* **16**, 240–246 (2006).
27. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
28. Friederici, A.D., Meyer, M. & von Cramon, D.Y. Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. *Brain Lang.* **74**, 289–300 (2000).
29. Canolty, R.T. *et al.* High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**, 1626–1628 (2006).
30. Lakatos, P. *et al.* An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.* **94**, 1904–1911 (2005).
31. Sirota, A., Csicsvari, J., Buhl, D. & Buzsáki, G. Communication between neocortex and hippocampus during sleep in rodents. *Proc. Natl. Acad. Sci. USA* **100**, 2065–2069 (2003).
32. Arnal, L.H. & Giraud, A.-L. Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* **16**, 390–398 (2012).
33. Poeppel, D., Idsardi, W.J. & van Wassenhove, V. Speech perception at the interface of neurobiology and linguistics. *Phil. Trans. R. Soc. Lond. B* **363**, 1071–1086 (2008).
34. Peña, M. & Melloni, L. Brain oscillations during spoken sentence processing. *J. Cogn. Neurosci.* **24**, 1149–1164 (2012).
35. Gross, J. *et al.* Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol.* **11**, e1001752 (2013).
36. Ding, N. & Simon, J.Z. Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.* **8**, 311 (2014).
37. Jackendoff, R. *Foundations of Language: Brain, Meaning, Grammar, Evolution* (Oxford University Press, 2002).
38. Hagoort, P. On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* **9**, 416–423 (2005).
39. Cutler, A., Dahan, D. & van Donselaar, W. Prosody in the comprehension of spoken language: a literature review. *Lang. Speech* **40**, 141–201 (1997).
40. Frazier, L., Carlson, K. & Clifton, C. Jr. Prosodic phrasing is central to language comprehension. *Trends Cogn. Sci.* **10**, 244–249 (2006).
41. Singer, W. & Gray, C.M. Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* **18**, 555–586 (1995).
42. Friederici, A.D. Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* **6**, 78–84 (2002).
43. Kutas, M. & Federmeier, K.D. Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn. Sci.* **4**, 463–470 (2000).
44. Neville, H., Nicol, J.L., Barss, A., Forster, K.I. & Garrett, M.F. Syntactically based sentence processing classes: evidence from event-related brain potentials. *J. Cogn. Neurosci.* **3**, 151–165 (1991).
45. Lau, E.F., Phillips, C. & Poeppel, D. A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* **9**, 920–933 (2008).
46. Halgren, E. *et al.* N400-like magnetoencephalography responses modulated by semantic context, word frequency and lexical class in sentences. *Neuroimage* **17**, 1101–1116 (2002).
47. Van Petten, C. & Kutas, M. Interactions between sentence context and word frequency in event-related brain potentials. *Mem. Cognit.* **18**, 380–393 (1990).
48. O'Connell, R.G., Dockree, P.M. & Kelly, S.P. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* **15**, 1729–1735 (2012).
49. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).
50. Nozaradan, S., Peretz, I., Missal, M. & Mouraux, A. Tagging the neuronal entrainment to beat and meter. *J. Neurosci.* **31**, 10234–10240 (2011).

ONLINE METHODS

Participants. 34 native listeners of Mandarin Chinese (19–36 years old, mean 25 years old; 13 male) and 13 native listeners of American English (22–46 years old, mean 26 years old; 6 male) participated in the study. All Chinese listeners received high school education in China and 26 of them also received college education in China. None of the English listeners understood Chinese. All participants were right-handed[51]. Five experiments were run for Chinese listeners and two experiments for English listeners. Each experiment included eight listeners (except that the AMS experiment involved five listeners) and each listener participated in at most two experiments. The number of listeners per experiment was chosen based on previous MEG experiments on neural tracking of continuous speech. The sample size for previous experiments was typically between three and 12 (refs. 15,16), and the basic phenomenon reported here was replicated in all the seven experiments of the study (N = 47 in total). The experimental procedures were approved by the New York University Institutional Review Board, and written informed consent was obtained from each participant before the experiment.

Stimuli I: Chinese materials. All Chinese materials were constructed based on an isochronous sequence of syllables. Even when the syllables were hierarchically grouped into linguistic constituents, no acoustic gaps were inserted between constituents. All syllables were synthesized independently using the Neospeech synthesizer (http://www.neospeech.com/, the male voice, Liang). The synthesized syllables were 75–354 ms in duration (mean duration 224 ms), and were adjusted to 250 ms by truncation or padding silence at the end. The last 25 ms of each syllable were smoothed by a cosine window.

Four-syllable sentences. 50 four-syllable sentences were constructed, in which the first two syllables formed a noun phrase and the last two syllables formed a verb phrase (**Supplementary Table 1**). The noun phrase could be composed of either a single two-syllable noun or a one-syllable adjective followed by a one-syllable noun. The verb phrase could be composed of either a two-syllable verb or a one-syllable verb followed by a one-syllable noun object. In a normal trial, ten sentences were sequentially played and no acoustic gaps were inserted between sentences (**Supplementary Fig. 1a**). Due to the lack of phrasal and sentential level prosodic cues, the sound intensity of the stimulus, characterized by the sound envelope, only fluctuates at the syllabic rate but not at the phrasal or sentential rate (**Supplementary Fig. 2**). An outlier trial was the same as a normal trial except that the verb phrases in two sentences were exchanged, creating two nonsense sentences with incompatible subjects and predicates (an example in English would be "new plans rub skin").

Four-syllable verb phrases. Two types of four-syllable verb phrases were created. Type I verb phrase contained a one-syllable verb followed by a three-syllable noun phrase, which could be a compound noun or an adjective + noun phrase (**Supplementary Fig. 1b** and **Supplementary Table 1**). Type II verb phrase contained a two-syllable verb followed by a two-syllable noun (**Supplementary Fig. 1c**, all phrases listed in **Supplementary Table 1**). 50 instances were created for each type of verb phrases. In a normal trial, ten phrases of the same type were sequentially presented. An outlier trial was the same as a normal trial except that the verbs in two phrases were exchanged, creating two nonsense verb phrases with incompatible verbs and objects (an example in English would be "drink a long walk").

Two-syllable phrases. The verb phrases (or the noun phrases) in the four-syllable sentences were presented in a sequence (**Supplementary Fig. 1d**). In a normal trial, 20 different phrases were played. In an outlier trial, one of the 20 phrases was replaced by two random syllables that did not constitute a sensible phrase.

Random syllabic sequences. The random syllabic sequences were generated based on the four-syllable sentences. Each four-syllable sentence was transformed into four random syllables using the following rule: the first syllable in the sentence was replaced by the first syllable of a randomly chosen sentence. The second syllable was replaced by the second syllable of another randomly chosen sentence and the same for the third and the fourth syllables. This way, if there were any consistent acoustic differences between the syllables at different positions in a sentence, those acoustic differences were preserved in the random syllabic sequences. Each normal trial contained 40 syllables. In outlier trials, four consecutive syllables were replaced by a Chinese idiom.

Backward syllabic sequences. In normal trials, ten four-syllable sentences were played but with all syllables being played backward in time. An outlier trial was the same as a normal trial except that four consecutive syllables at a random position were replaced by four random syllables that were not reversed in time.

Four-syllable idioms. 50 common four-syllable idioms were selected (**Supplementary Table 1**), in which the first two syllables formed a noun phrase and the last two syllables formed a verb phrase. In a normal trial ten sentences were played. An outlier trial was the same as a normal trial except that the noun phrases in two idioms were exchanged, creating two nonexistent and semantically nonsensical idioms.

Sentences with variable duration and syntactic structures. The sentence duration was varied between four and eight syllables. 40 sentences were constructed for each duration, resulting in a total of 200 sentences (listed in **Supplementary Table 1**). All 200 sentences were intermixed. In a normal trial, ten different sentences were sequentially played without inserting any acoustic gap in between sentences. In an outlier trial, one of the ten sentences was replaced by a syntactically correct but semantically anomalous sentence. Examples of nonsense sentences, translated into English, included "ancient history is drinking tea" and "take part in his portable hard drive".

Sentences with variable NP durations. All sentences consisted of a noun phrase followed by a verb phrase (**Supplementary Table 1**). The noun phrase had three syllables for half of the sentences (N = 45) and four syllables for the other half. A three-syllable noun phrase was followed by either a four-syllable verb phrase (N = 20) or a five-syllable verb phrase (N = 25). A four-syllable noun phrase was followed by a three-syllable verb phrase (N = 20) or a four-syllable verb phrase (N = 25). Sentences with different noun phrase durations and verb phrase durations were intermixed. In a normal trial 10 different sentences were played sequentially, without inserting any acoustic gap between phrases or sentences. In an outlier trial one sentence was replaced by a sentence with the same syntactic structure but that was semantically anomalous.

AMS. Five sets of AMS were created. Each sentence consisted of three components, C1, C2 and C3. Each component (C1, C2 or C3) was independently chosen from three candidate syllables with equal probability. The grammar of the AMS is illustrated in **Supplementary Figure 4a**. In the experiments, sentences were played sequentially without any gap between sentences. Since all components were chosen independently and each component was chosen from three syllables with equal probability, all components were equally predictable regardless of its position in a sequence. In other words, $P(C1) = P(C2) = P(C3) = P(C2|C1) = P(C3|C2) = P(C1|C3) = 1/3$.

All Chinese syllables were synthesized independently and adjusted to 300 ms by truncation or padding silence at the end. In each trial, 60 sentences were played and no additional gap was inserted between sentences. Therefore, the syllables were played at a constant rate of 3.33 Hz and the sentences were played at a constant rate of 1.11 Hz. To make sure that neural encoding of the AMS was not confounded by acoustic properties of a particular set of syllables, five sets of AMS were created (**Supplementary Table 1**). No meaningful Chinese expressions are embedded in the AMS sequences.

Stimuli II: English materials. All English materials were synthesized using the MacinTalk Synthesizer (male voice Alex, in Mac OS X 10.7.5).

Four-syllable English sentences. 60 four-syllable English sentences were constructed (**Supplementary Table 1**), and each syllable was a monosyllabic word. All sentences had the same syntactic structure: adjective/pronoun + noun + verb + noun. Each syllable was synthesized independently, and all the synthesized syllables (250–347 ms in duration) were adjusted to 320 ms by padding silence at the end or truncation. The offset of each syllable was smoothed by a 25-ms cosine window. In each trial, 12 sentences were presented without any acoustic gap between them. In an outlier trial, 3 consecutive words from a random position were replaced by three random words so that the corresponding sentence(s) became ungrammatical.

Shuffled sequences. Shuffled sequences were constructed as an unintelligible sound sequence that preserved the acoustic properties of the sentence sequences. All syllables in the four-syllable English sentences were segmented into five overlapping slices. Each slice was 72 ms in duration and overlapped with neighboring slices for 10 ms. The first 10 ms and the last 10 ms of each slice was smoothed by a linear ramp, except for the onset of the first slice and the offset of the last slice.

doi:10.1038/nn.4186

A shuffled 'sentence' was constructed by shuffling all slices at the same position across the four-syllable sentences. For example, the first slice of the first syllable in a given sentence was replaced by the first slice of the first syllable in a different randomly chosen sentence. For another example, the third slice of the fourth syllable in one sentence was replaced by the third slice of the fourth syllable in another randomly chosen sentence. In a normal trial, 12 different shuffled sentences were played sequentially, resulting in a trial that had the same duration as a trial of four-syllable English sentences. In an outlier trial, four consecutive shuffled syllables were replaced by four randomly chosen English words that did not construct a sentence.

Markovian sentences. The pronunciation of an English syllable strongly depends on its neighbors. To simulate more natural English, we also synthesized English sentences based on an isochronous multi-syllabic 'acoustic chunk'. Every sentence was divided into three acoustic chunks that were roughly equal in duration. Each acoustic chunk consisted of 1–2 monosyllabic or bisyllabic words and was synthesized as a whole, independently of neighboring acoustic chunks. All synthesized acoustic chunks (250–364 ms in duration) were adjusted to 350 ms by truncation or padding silence at the end. The offset of each chunk was smoothed by a 25-ms cosine window.

Two types of Markov chain sentences were generated based on isochronous sequences of acoustic chunks. In one type of Markovian sentences, called the constant predictability sentences, each acoustic chunk was equally predictable based on the preceding chunk, regardless of its position within a sentence. The constant predictability sentences were generated based on the grammar specified in **Figure 3a** and **Supplementary Figure 1e**. Listeners were familiarized with the grammar and were able to write down the full grammar table before participating in the experiment. In each trial, ten sentences were separately generated based on the grammar and sequentially presented without any acoustic gap between them.

The other type of Markovian sentences, called the predictable sentences, consisted of a finite number of sentences (N = 25, **Supplementary Table 1**) that were extensively repeated (11–12 times) in a ~7-min block. In these sentences, the second and the third acoustic chunks were uniquely determined by the first chunk. In each trial, ten different sentences were played sequentially without any acoustic gap between them.

Acoustic analysis. The intensity fluctuation of the sound stimulus is characterized by its temporal envelope. To extract the temporal envelope, the sound signal is first half-wave rectified and then downsampled to 200 Hz. The Discrete Fourier Transform of the temporal envelope (without any windowing) is shown in **Figure 1** and **Supplementary Figure 2**.

Experimental procedures. Seven experiments were run. Experiment 1–4 involved Chinese listeners listening to Chinese materials, experiment 5 involved English listeners listening to Chinese materials, and experiment 6 involved English listeners listening to English materials. Experiment 7 involved Chinese listeners listening to AMS.

In all experiments except for experiment 5, listeners were instructed to detect outlier trials. At the end of each trial, listeners had to report whether it was a normal trial or an outlier trial via button press. Following the button press, the next trial was played after a delay randomized between 800 and 1,400 ms. In experiment 5, listeners performed a syllable counting task described below. Behavioral results are reported in **Supplementary Table 2**.

Experiment 1. Four-syllable Chinese sentences, four-syllable idioms, random syllabic sequences and backward syllabic sequences were presented in separate blocks. The order of the blocks was counter balanced across listeners. Listeners took breaks between blocks. In each block, 20 normal trials and ten outlier trials were intermixed and presented in a random order.

Experiment 2. Four-syllable sentences, type I four-syllable verb phrases, type II four-syllable verb phrases, two-syllable noun phrases, and two-syllable verb phrases were presented in separate blocks. The order of the blocks was counter balanced across listeners. Listeners took breaks between blocks. In each block, 20 normal trials and five outlier trials were intermixed and presented in a random order.

Experiment 3. Sentences with variable durations and syntactic structures, as described above, were played in an intermixed order. Listeners took a break every 25 trials. In total, 80 normal trials and 20 outlier trials were presented.

Experiment 4. Sentences with variable NP durations, as described above, were presented in a single block, counterbalanced with three other blocks that presented language materials not analyzed here. In that block, 27 normal trials and seven outlier trials were presented. The other three blocks presented other language materials not analyzed here. The order of the blocks was counterbalanced across listeners.

Experiment 5. Trials consisting of four-syllable sentences, four-syllable idioms, random syllabic sequences, and backward syllabic sequences were intermixed and presented in a random order. 20 normal trials for each type of materials were presented. In each trial, the last 1 or 2 syllable was removed, each with 50% probability. Listeners were instructed to count the number of syllables in each trial in a cyclic way: 1, 2, 3, 4, 1, 2, 3, 4, 1, 2 … The final count could only be 2 or 3 and the listeners had to report whether it was 2 or 3 at the end of each trial via button press.

Experiment 6. Four-syllable English sentences, shuffled sequences, constant predictability Markovian sentences, and predictable Markovian sentences were presented in separate blocks. The order of the blocks was counterbalanced across listeners. Listeners took breaks between blocks. In each block, 22 normal trials and 8 outlier trials were intermixed and presented in a random order.

Experiment 7. The experiment involved the AMS and was divided into two sessions. In the first session, ten trials were presented (two trials from each AMS set; see the upper row in **Supplementary Fig. 4b**). In each trial, the last syllable was removed with 50% probability. The listeners were told that the stimulus was only a sequence of random syllables. They were asked to count the number of syllables in a cyclic way: 1, 2, 1, 2, 1, 2 … and report whether the final count was 1 or 2 at the end of each trial via button press. Since each trial contained 179 or 180 rapidly presented syllables, the listeners were not able to count accurately (mean performance 52 ± 9.7%, not significantly above chance, $P > 0.8$, t test). However, the listeners were asked to follow the rhythm and keep counting even when they lost count. After the first session of the experiment was finished, the listeners were told about that the general structure of the AMS and examples were given based on real Chinese sentences. In the second session of the experiment, the listeners had to learn the 5 sets of AMS separately (lower row, **Supplementary Fig. 4b**). For each set of the AMS, during training, the listeners listened to 20 sentences from the AMS set in a sequence, with a 300-ms gap being inserted between sentences to facilitate learning. Then, the listeners listened to two trials of sentences from the same AMS set, which they also listened to in the first session (shown by symbol S in **Supplementary Fig. 4b**). They had to do the same cyclical counting task. However, they were told that the last count was 1 if the last sentence was incomplete and the last count was 2 if the last sentence was complete (mean performance 82 ± 8.0%, significantly above chance $P < 0.2$, t test). At the end of the two trials, the listeners had to report the grammar of the AMS, i.e. which 3 syllables could be the first syllable of a sentence, which three syllables could be the middle one, and which three syllables could be the last one. The grammatical roles of 77 ± 7.6% (mean ± s.e. across subjects) syllables were reported correctly.

Neural recordings. Cortical neuromagnetic activity was recorded using a 157-channel whole-head MEG system (KIT) in a magnetically shielded room. The MEG signals were sampled at 1 kHz, with a 200-Hz low-pass filter and a 60-Hz notch filter applied online and a 0.5-Hz high-pass filter applied offline (time delay compensated). The environmental magnetic field was recorded using three reference sensors and regressed out from the MEG signals using time-shifted PCA[52]. Then, the MEG responses were further denoised using the blind source separation technique, Denoising Source Separation (DSS)[53]. The MEG responses were decomposed into DSS components using a set of linear spatial filters, and the first 6 DSS components were retained for analysis and transformed back to the sensor space. The DSS decomposes multi-channel MEG recordings to extract neural response components that are consistent over trials and has been demonstrated to be effective in denoising cortical responses to connected speech[18,54,55]. The DSS was applied to more accurately estimate the strength of neural activity phase-locked to the stimulus. Even when the DSS spatial filtering process was omitted, for the r.m.s. response over all MEG sensors, the sentential, phrasal, and syllabic responses in **Figure 1** were still statistically significant ($P < 0.001$, bootstrap).

Data analysis. Only the MEG responses to normal trials were analyzed.

Frequency domain analysis. In experiments 1, 2, 5 and 6, the linguistic structures of different hierarchies were presented at unique and constant rates

and neural tracking of those linguistic structures was analyzed in the frequency domain. For each trial, to avoid the transient response to the acoustic onset of each trial, the neural recordings were analyzed in a time window between the onset of the second sentence (or the fifth syllable if the stimulus contained no sentential structure) and the end of the trial. The single-trial responses were transformed into the frequency domain using the discrete Fourier transform (DFT). For all Chinese materials and the artificial Markovian language materials, nine sentences were analyzed in each trial, resulting in a frequency resolution of 1/9 of the sentential rate (~0.11 Hz). For the English sentences and the shuffled sequences, the trials were longer and the duration equivalent to 44 English syllables was analyzed, resulting in a frequency resolution of 1/44 of the syllabic rate, that is, 0.071 Hz.

The response topography (**Fig. 1c**) showed the power of the DFT coefficients at a given frequency and hemispheric lateralization was calculated by averaging the response power over the sensors in each hemisphere ($N = 54$).

Given that the properties of the neural responses to linguistic structures and background neural activity might vary in different frequency bands, to treat each frequency band equally, a separate spatial filter was designed for every frequency bin in the DFT output to optimally estimate the response strength. The linear spatial filter was the DSS filter[56]. The output of the DSS filter was a weighted summation over all MEG sensors, and the weights were optimized to extract neural activity consistent over trials. In brief, if the DFT of the MEG response averaged over trials is $\boldsymbol{X}(f)$ and the autocorrelation matrix of single-trial MEG recordings is $\boldsymbol{R}(f)$, the spatial filter is $\boldsymbol{w} = \boldsymbol{R}^{-1}(f)\boldsymbol{X}(f)$ (see the appendix of ref. 56). The spatial filter $\boldsymbol{w}$ is an 157×1 vector (for the 157 sensors), the same size as $\boldsymbol{X}(f)$, and $\boldsymbol{R}(f)$ is a 157×157 matrix. The spatial filter could be viewed as a virtual sensor that was optimized to record phase-locked neural activity at each frequency. Power of the scalar output of the spatial filter, $|\boldsymbol{X}^{\mathrm{T}}(f)\boldsymbol{R}^{-1}(f)\boldsymbol{X}(f)|^2$, was the power spectral density shown in the figures.

Time domain analysis. The response to each sentence was baseline corrected based on the 100-ms period preceding the sentence onset, for each sensor. To remove the neural response to the 4-Hz isochronous syllabic rhythm and focus on the neural tracking of sentential/phrasal structures, we low-pass filtered the neural response waveforms using a 0.5-s duration linear phase FIR filter (cut-off 3.5 Hz). The filter delay was compensated by filtering the neural signals twice, once forward and once backward. When the response power at 4 Hz was extracted separately by a Fourier analysis, it does not significantly change as a function of sentence duration ($P > 0.19$, one-way ANOVA). The r.m.s. of the MEG responses was calculated as the sum of response power (that is, square of the MEG response) of all sensors, and the r.m.s. response was further low-pass filtered by a 0.5-s duration linear phase FIR filter (cut-off 3.5 Hz, delay compensated).

A linear decoder was built to decode the duration of sentences. In the decoding analysis, the multi-channel MEG responses were compressed to a single channel, i.e. the first DSS component, and the decoder solely relied on the time course of the neural response. A 2.25-s response epoch was extracted for each sentence, starting from the sentence onset. A leave-one-out cross-validation procedure was employed to evaluate the decoder's performance. Each time, the response to one sentence was used as the testing response, and the responses to all other sentences were treated as the training set. The training signals were averaged for sentences of the same duration, creating a template for the response time course for each sentence duration. The testing response was correlated with all the templates and the category of the most correlated templates was the decoder's output. For example, if the testing response was most correlated with the template for 5-syllable sentences, the decoder's output would be that the testing response was generated by a five-syllable sentence.

Statistical analysis and significance tests. For spectral peaks (**Figs. 1** and **2**), a one-tailed paired *t* test was used to test if the neural response in a frequency bin was significantly stronger than the average of the neighboring four frequency bins (two bins on each side). Such a test was applied to all frequency bins below 5 Hz, and a FDR correction for multiple comparisons was applied. Except for the analysis of the spectral peaks, two-tailed *t* tests were applied. For all the *t* tests applied in this study, data from the two conditions had comparable variance and showed no clear deviation from the normal distribution when checking the histograms. If the *t* test was replaced by a bias-corrected and accelerated bootstrap, all results remained significant.

In **Figure 4**, the s.e.m. over subjects was calculated using bias-corrected and accelerated bootstrap[57]. In the bootstrap procedure, all the subjects were resampled with replacement 2,000 times. The s.d. of the resampled results was taken as the s.e.m. In **Figure 4d**, the statistical difference between the two curves, that is, the three-syllable NP condition and the four-syllable NP condition, was also tested using bootstrap. For each subject, the difference between the responses in these two conditions was taken. At each time point, the response difference was resampled with replacement 2,000 times across the eight subjects, and percentage of the resampled differences being larger or smaller than 0 (the smaller value) was taken as the significance level. A FDR correction was applied to the bootstrap results.

Code availability. The computer code used for the MEG analyses is available upon request.

Neural Source Localization Using ECoG. *ECoG participants.* ECoG recordings were obtained from five patients (three female; average 33.6 years old, range 19–42 years old) diagnosed with pharmaco-resistant epilepsy and undergoing clinically motivated subdural electrode recordings at the New York University Langone Medical Center. Patients provided informed consent before participating in the study in accordance with the Institutional Review Board at the New York University Langone Medical Center. Three patients were right-handed, two were left-handed. All patients were native English speakers (one of them was a bilingual Spanish/English speaker), and all patients were left-hemisphere dominant for language.

ECoG recordings. Participants were implanted with 96–179 platinum-iridium clinical subdural grid or strip electrodes (three patients with a left-hemisphere implant and two patients with a right hemisphere implant, additional depth electrodes implanted for some patients but not analyzed). The electrode locations per patient are shown in **Supplementary Figure 5**. Electrode localization followed previously described procedures[58]. In brief, for each patient we obtained pre-op and post-op T1-weighted MRIs which were co-registered with each other and normalized to a MNI-152 template, allowing the extraction of the electrode location in MNI space.

The ECoG signals were recorded with a Nicolet clinical amplifier at a sampling rate of 512 Hz. The ECoG recordings were re-referenced to the grand average over all electrodes (after removing artifact-laden or noisy channels). Electrodes from different subjects were pooled per hemisphere, resulting in 385/261 electrodes in the left/right hemispheres. High gamma activity was extracted by high-pass filtering the ECoG signal above 70 Hz (with additional notch filters at 120 and 180 Hz). The energy envelope of high gamma activity was extracted by taking the square of high-gamma response waveform.

ECoG procedures. Participants performed the same task as healthy subjects in the MEG (**Fig. 2e**,**f**). In brief, they listened to a set of English sentences and control stimuli in the first and second block. The control stimulus, that is, the shuffled sequences, preserves the syllabic-level acoustic rhythm of English sentences but contain no hierarchical linguistic structure. The procedure was the same as the MEG experiment, except for a familiarization session in which the subjects listened to individual sentences with visual feedback. 60 trials of sentences and control stimuli were played. The ECoG data from each electrode was analyzed separately and converted to the frequency domain via DFT (frequency resolution 0.071 Hz).

A significant response at the syllabic, phrasal or sentential rate was reported if the response power at the target frequency was stronger than the response power averaged over neighboring frequency bins (0.5-Hz range above and below the target frequency). The significance level for each electrode was first determined based on a bootstrap procedure that randomly sampled the 60 trials 1,000 times and then underwent FDR correction for multiple comparisons across all electrodes in the same hemisphere.

A **Supplementary Methods Checklist** is available.

51. Oldfield, R.C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
52. de Cheveigné, A. & Simon, J.Z. Denoising based on time-shift PCA. *J. Neurosci. Methods* **165**, 297–305 (2007).
53. de Cheveigné, A. & Simon, J.Z. Denoising based on spatial filtering. *J. Neurosci. Methods* **171**, 331–339 (2008).
54. Ding, N. & Simon, J.Z. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.* **33**, 5728–5735 (2013).
55. Ding, N. & Simon, J.Z. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* **107**, 78–89 (2012).
56. Wang, Y. *et al.* Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. *J. Neurophysiol.* **107**, 2033–2041 (2012).
57. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* (CRC press, 1993).
58. Yang, A.I. *et al.* Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage* **63**, 157–165 (2012).

doi:10.1038/nn.4186