Contents

Supplementary Information.

Neural tracking of linguistic structures.

Experiment 1 served as the experimental baseline, which was conducted to replicate and extend the cortical tracking effect originally found by Ding et al. (2016). In this experiment, Dutch participants listened to three types of Dutch syllable sequences, which were disyllabic noun sequences (T1), random syllable sequences (T2) and backward played random syllable sequences (T3), respectively (see **Supplementary Fig. 2a** and **Methods**). The sequences were aurally presented at a rhythm of 4 syllables per second (4 Hz) in a random order, and the corresponding neural electromagnetic responses were recorded. The power spectrum (spatially filtered) was separately extracted for each of the three conditions. Statistical analysis (paired sample t-test, Bonferroni corrected) on the power of T1 sequences suggested a significantly stronger response at 2 Hz ($t(13) = 10.13$, $p < 7.73e\text{-}08$, **Supplementary Fig. 2b**) and 4 Hz ($t(13) = 5.19$, $p < 8.62e\text{-}05$, **Supplementary Fig. 2b**) compared to their corresponding neighboring bins. In contrast, only a significant 4 Hz peak occurred when participants listened to T2 ($t(13) = 6.44$, $p < 1.09e\text{-}05$, **Supplementary Fig. 2b**) and T3 ($t(13) = 5.88$, $p < 2.69e\text{-}05$, **Supplementary Fig. 2b**) sequences. The topographical distributions in **Supplementary Fig. 2b** depict the weight of each sensor in extracting the optimized time series at 2 Hz and 4 Hz (for T1 sequences only), with bigger red dots indicating higher absolute weights (for details see **Methods**).

To explore the cortical origin of the power effect, source reconstructions were conducted at 2 and 4 Hz (T1 sequences, for details see **Methods**). A cluster-based permutation test at 4 Hz (the syllables' rhythm) indicated that the T1 sequences held

stronger source power compared to the baseline (T1 sequences, t_s = 3.62e+06, p_c < 0.01; see **Supplementary Fig. 2d**), and the effect was most pronounced bilaterally in the frontal (IFG), temporal (ITG, MTG, STG, INS) and central (PrG) regions, along with the left SFG, MFG, IPL and PoG (see **Methods** for the full names of the abbreviations). To estimate the magnitude of the effect, a paired sample t-test was used to compare the average power within the cluster between the target and baseline conditions. As expected, a robust 4 Hz source power corresponding to the target condition was observed (T1 sequences, t (13) = 4.10, p < 0.0014, see lower panel of **Supplementary Fig. 2d**). The same estimation pipeline was applied on the power response at 2 Hz (the words' rhythm). A significant source power was initially identified (for T1 sequences, t_s = 2.34e+06, p_c < 0.05, see **Supplementary Fig. 2c**), which was bilaterally distributed in the frontal (MFG) and temporal (STG, INS) regions, together with the areas that had a left hemispheric dominance including MTG, ITG, IFG and PrG. Further estimation on the magnitude of the effect suggested that the source power of the target condition was robustly higher than the baseline (t (13) = 4.78, p < 3.56e-04, see the lower panel of **Supplementary Fig. 2c**).

Consistent with previous studies (Ding et al., 2016; Gui et al., 2020; Jin et al., 2018; Kaufeld et al., 2020; Lu et al., 2023; Martin & Doumas, 2017; Ten Oever et al., 2022), our sensor space results indicated that the power of neural oscillations can simultaneously tracked the linguistic structures (word and syllables) at different time scales. The results suggested that the power activity was an effective neural readout to reflect speech hierarchy. Furthermore, the source localizations identified for words (2 Hz) and syllables (4 Hz) showed a strong left hemispheric bias and overlapped with the typical regions for

speech processing (Giraud & Poeppel, 2012; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007). The source distributions were expected since the participants were listening to the sequences in their native language, and therefore, the cortical areas associated with language related processing should be involved.

To assess the role of phase activity and distinguish it from the power in representing speech hierarchy, inter-trial phase coherence (ITPC) was calculated at all frequency bins (for details see **Methods**). Statistical analysis (paired sample t-test, FDR corrected) on the T1 sequences indicated that the strength of the phase coherence at 2 Hz ($t(13) = 11.05$, $p < 2.76e-08$, the left panel of **Supplementary Fig. 2e**) and 4 Hz ($t(13) = 11.98$, $p < 1.06e-08$, the left panel of **Supplementary Fig. 2e**) were significantly higher than the baseline (the orange line in **Supplementary Fig. 2e**). In comparison, a robust phase synchronization was found only at 4 Hz for T2 ($t(13) = 14.55$, $p < 1.00e-09$, the middle panel of **Supplementary Fig. 2e**) and T3 sequences ($t(13) = 11.61$, $p < 1.55e-08$, the right panel of **Supplementary Fig. 2e**). Note that we found no evidence to suggest that the angles of the phase coherence were consistent among participants. In other words, while each participant's phase activity clustered around a specific angle across trials, these angles were not statistically consistent across participants (see the same type of transparent lines, such as dotted lines, in **Supplementary Fig. 2e**).

To further estimate the source origin of the phase coherence and check its temporal evolution, source reconstructions at multiple frequencies were conducted on T1 sequences (for details see **Methods**). Statistical analysis (cluster-based permutation test) on the averaged phase coherence (averaged over a 3-second window from 2 to 4 seconds after the audio onset) indicated that the source phase coherence at 4 Hz was significantly

higher than the average of those at its neighbor frequency bins (T1 sequences, t_s = 7.86e+06, p_c < 0.002, see **Supplementary Fig. 2g**). And the effect was most pronounced bilaterally at the frontal (IFG, MFG, SFG, OrG), temporal (ITG, MTG, STG, INS) and central areas (IPL, SPL, PrG, PoG). The lower panel of **Supplementary Fig. 2g** shows the temporal evolution of the averaged phase coherence for both conditions. By applying the same estimation pipeline, we found a significant 2 Hz phase coherence as well (T1 sequences, t_s = 7.26e+06, p_c < 0.002, see **Supplementary Fig. 2f**). And the effect was largely localized bilaterally in the frontal (MFG, SFG, OrG), temporal (ITG, STG) and central (PrG, PoG) areas, along with the left MTG and the right IPL. The temporal evolution of the averaged phase coherence at 2 Hz for both conditions were shown at the lower panel of **Supplementary Fig. 2f**.

The sensor level results indicated that phase activity was associated with representing different types of linguistic units and their hierarchical relations during speech comprehension. The periodicities of linguistic structures were reflected by both phase and power suggested that the two neural readouts could be associated with different processes in building hierarchical structures. In addition, the source distribution for phase was quite different from that for power (e.g., in terms of the involved cortical regions), which again suggested the potential difference in roles between the two neural measures.

Experiment 1 replicated the original cortical tracking effect (Ding et al., 2016) and then explored the role of phase activity in representing hierarchical structures. In addition, the cortical origins of the two types of neural readouts were estimated, which provided in-depth information to inspect the phenomenon. As discussed above (see

Introduction), the observed effect is driven by prosodic, statistical, and structural linguistic cues. In addition to the co-extension of linguistic and statistical information (the TP between words in T1 sequences was 1/10, see **Methods**), there were measurable prosodic cues to indicate the words structures (2 Hz units) in T1 sequences (2 Hz peak in acoustic spectrum, see **Methods** and **Supplementary Fig. 6g**). The source localization for power overlapped with the cortical regions related to language processing, however, given the existence of the prosodic and TP differences in conditions, it is a bit difficult to argue that the observed cortical network was predominantly driven by linguistic cues. In addition, the 2 Hz peak here was relatively higher than the one showed in the original study (Ding et al., 2016), which could be induced by the accessibility of multiple kinds of cues (e.g., prosodic, linguistic and statistical). It was highly likely that the different types of cues all together contributed to the 2 Hz peak here, therefore, our results did not eliminate the possibility that linguistic information alone can elicit the tracking phenomenon.

Neural tracking of statistical structures.

The results of **Experiment 1** demonstrated that our experimental paradigm effectively elicited the tracking effect when multiple types of cues were available. And, as expected, neural phase activity was shown to be involved in representing hierarchical structures. In **Experiment 2**, we removed linguistic cues and to at a large extent attenuated the influence of physical (e.g., prosodic) cues (see **Methods** and **Supplementary Fig. 6h**) to see how statistical information alone (i.e., the TP between words was 1/10, see **Methods**) would reshape the tracking phenomenon (in time, frequency and space). In this experiment, Dutch participants listened to the same three

types of sequences (T1, T2 and T3), and the experimental procedure and stimuli's manipulations were identical to **Experiment 1** except that the stimuli were in Mandarin Chinese rather than Dutch (linguistic knowledge removed, see **Supplementary Fig. 3a**).

Statistical analysis (paired sample t-test, FDR corrected) on the power spectrum of T1 sequences indicated that the peaks at 2 Hz ($t(13) = 4.25$, $p < 4.69e\text{-}04$, see **Supplementary Fig. 3b**) and 4 Hz ($t(13) = 8.89$, $p < 3.49e\text{-}07$, see **Supplementary Fig. 3b**) were significantly higher than their corresponding neighbor bins (the topographies show the weights of the sensors). In contrast, only a 4 Hz peak occurred for T2 ($t(13) = 8.75$, $p < 4.14e\text{-}07$, see **Supplementary Fig. 3b**) and T3 ($t(13) = 6.53$, $p < 9.47e\text{-}06$, see **Supplementary Fig. 3b**) sequences.

Further source level estimations suggested that the source power at 4 Hz (syllables) for T1 sequences was significantly higher than that for the baseline condition (cluster-based permutation test, $t_s = 3.31e+06$, $p_c < 0.01$). And the effect was most pronounced bilaterally in the frontal (IFG, MFG), temporal (ITG, MTG, STG, INS) and central (PrG, PoG) areas, along with the right IPL (see the upper panel of **Supplementary Fig. 3d**). The estimation on effect size suggested that the averaged power within the cluster was robustly stronger for the target condition (T1 sequences) than the baseline condition ($t(13) = 3.50$, $p < 0.0038$, see the lower panel of **Supplementary Fig. 3d**). Similarly, statistical analysis on the 2 Hz source power indicated that the intensity of the target condition (T1 sequences) was significantly higher than that of the baseline condition (cluster-based permutation test, $t_s = 3.57e+06$, $p_c < 0.01$). And the effect was largely distributed bilaterally at the frontal (IFG, MFG, SFG, OrG) and temporal (ITG, MTG, STG, INS) areas (see the upper panel of **Supplementary Fig. 3c**), with a slight bias towards

the right hemisphere in terms of activation strength (t-values) and the extent of activation within specific cortical regions (e.g., ITG). Additional analysis on the magnitude of the effect validated its robustness (paired sample t-test, $t(13) = 6.19$, $p < 3.25e\text{-}05$, see the lower panel of **Supplementary Fig. 3c**).

In **Experiment 2**, the prosodic (see **Methods** and **Supplementary Fig. 6h**) and linguistic cues (Dutch participants listened to Mandarin Chinese) for word recognition were unavailable, a robust power response occurred at 2 Hz (words) when the participants listened to T1 sequences suggested that statistical information alone can induce the tracking effect. In addition, when considering the results of the original study (Ding et al., 2016), which showed that linguistic information alone could induce the effect, along with the sensor-level power results from **Experiment 1** (where multiple types of cues were available) and **Experiment 2** (where only statistical cues were available), it is reasonable to conclude that hierarchical representation can be elicited through either one type or multiple types of perceptual cues.

However, we noticed that the cortical origin of the power response when multiple kinds of cues were available (see **Supplementary Fig. 2c** and **Supplementary Fig. 2d**) was quite different from that when only statistical information was provided (see **Supplementary Fig. 3c** and **Supplementary Fig. 3d**). The source variations between the two situations suggested that the cortical regions that underpinned the utilization of different types of cues were different. Though the differences in stimuli (Dutch vs Mandarin Chinese) might have contributed to the source variations, explaining the effect by the availability of structural cues seemed more plausible, since the source power results in the following experiments indicated that the cortical origins were fairly similar when

only one type of cue (statistical) was provided given the languages of the stimuli differed across experiments (see **Supplementary Fig. 4c** and **Supplementary Fig. 5c**). Moreover, the 2 Hz peak when multiple cues were available was higher than that when only the statistical cues were provided (paired sample t-test, $p < 1.00e-03$), which potentially suggested the accumulated effect of cues for structure building. In sum, the results up to here indicated that hierarchical representation reflected by the neural power response can be elicited through one or multiple cues, and the strength of the peak at one frequency (e.g., 2 Hz) and its corresponding source origins might reflect the availability and the role of the cues for building hierarchy.

To examine how phase activity was involved in the construction of hierarchical structures when only statistical cues was provided, phase coherence (ITPC, see **Methods**) was estimated at both the sensor and source levels. Statistical comparisons (paired sample t-test, FDR corrected) at sensor level were first conducted on the phase coherence corresponding to T1 sequences. As expected, the analyses indicated that the strength of the phase synchronization across trials was significantly higher at 2 Hz ($t(13) = 5.75$, $p < 3.34e-05$, see left panel of **Supplementary Fig. 3e**) and 4 Hz ($t(13) = 17.15$, $p < 1.31e-10$, see left panel of **Supplementary Fig. 3e**) compared to the baseline. In contrast, only a significant 4 Hz phase coherence occurred for T2 ($t(13) = 12.74$, $p < 5.04e-09$, see middle panel of **Supplementary Fig. 3e**) and T3 ($t(13) = 10.04$, $p < 8.55e-08$, see right panel of **Supplementary Fig. 3e**) sequences. Similar to **Experiment 1**, no statistical evidence was found to support that the phase coherence effect was driven by one specific phase angle across participants in any of the three situations (see **Supplementary Fig. 3e**).

Further source estimations for the phase coherence were applied on T1 sequences. Statistical comparisons (cluster-based permutation test) suggested that the strength of the phase coherence at 4 Hz was significantly higher than at neighboring frequency bins (T1 sequences, t_s = 5.61e+06, p_c < 0.002, for details see **Methods**). And the effect was most pronounced bilaterally in the frontal (IFG, MFG, SFG, OrG), temporal (ITG, MTG, pSTS) and central regions (PoG, PrG), along with the right IPL, right STG, and left SPL (see the upper panel of **Supplementary Fig. 3g**). The lower panel of **Supplementary Fig. 3g** shows the averaged phase coherence within the cluster for both conditions. Similarly, statistical comparisons (cluster-based permutation test) suggested that the phase coherence at 2 Hz was robustly stronger compared to the baseline (T1 sequences, t_s = 6.31e+06, p_c < 0.002). And the effect was localized bilaterally in the frontal (IFG, MFG, SFG) and temporal areas (ITG, MTG, STG, pSTS), together with left PrG, left PoG and right SPL (see the upper panel of **Supplementary Fig. 3f**). The temporal dynamics in the lower panel of **Supplementary Fig. 3f** represent the averaged phase coherence within the cluster for both conditions.

Apparently, the results indicated that the representation of hierarchical structures was associated with phase activity, even when only statistical information was provided for building structures. Consistent with **Experiment 1**, the cortical origins of the phase response (see **Supplementary Fig. 3f** and **Supplementary Fig. 3g**) in **Experiment 2** differed from those of the power response (see **Supplementary Fig. 3c** and **Supplementary Fig. 3d**). The findings suggested that the phase and power measures are potentially linked to distinct processes involved in the hierarchical representation, regardless how many types of cues were available. Later result sections discussed the roles

associated with these two types of neural measurements (see the theoretical model) and what the underlying source distributions represent (see the results of connectivity).

Neural tracking of linguistic and statistical structures.

Previous experiments demonstrated that hierarchical representations can be elicited by either one type (**Experiment 2**) or multiple types of structural cues (**Experiment 1**) where the stimuli contained two types of units (i.e., words and syllables). To generalize the tracking effect, it is important to examine whether the hierarchical representation still occurs when multiple layers are present. To asses that, we generated a type of sequence (noun pairs, 1 Hz, 4 syllables per second) that was layered on top of the word rate (2 Hz) and violates grammatical rules in Dutch. By doing so, we aimed to check if the brain would simultaneously represent the units at different levels (i.e., noun pairs, words and syllables) and to investigate how the brain would handle a structure that was statistically associated but violated grammatical expectation (for details see **Methods**).

This section involved two experiments. In the first experiment (**Experiment 3**), we trained the participants on the noun pairs (1 Hz structures) from sequences that were varied in durations (1, 2 or 3 seconds, see **Supplementary Fig. 7**). These noun pairs were formed by combining two singular nouns, which violated grammatical rules in Dutch. During the training, the TP between any two noun pairs was controlled to be 1/25, which served as the statistical cue for learning (for details see **Methods**). After the training, **Experiment 4** was conducted, in which the Dutch participants listened to the same three types of sequences (T1, T2 and T3, see **Supplementary Fig. 4a**) as in **Experiments 1** and **2**, except that the sequences were constructed by using the trained stimuli from **Experiment 3**. Specifically, in each T1 sequence, each pair of words (without

replacement) formed a noun pair (1 Hz) composed of two singular nouns (2 Hz). Similar paradigms were used to explore the role of statistical information in extracting artificially constructed units (Henin et al., 2021; Saffran et al., 1999). Note that the analyses were conducted solely on the neural activities recorded from **Experiment 4**.

The sensor level power response was initially analyzed. Statistical comparisons on T1 sequences (paired sample t-test, FDR corrected) indicated that the strength of the induced power at 1 Hz ($t(13) = 4.94$, $p < 1.33e-04$, see **Supplementary Fig. 4b**), 2 Hz ($t(13) = 12.26$, $p < 8.00e-09$, see **Supplementary Fig. 4b**) and 4 Hz ($t(13) = 5.10$, $p < 1.01e-04$, see **Supplementary Fig. 4b**) were significantly higher than their corresponding neighbor bins. However, as expected, only a significant 4 Hz response was observed when participants listened to T2 ($t(13) = 6.07$, $p < 1.98e-05$, see **Supplementary Fig. 4b**) and T3 ($t(13) = 3.80$, $p < 0.001$, see **Supplementary Fig. 4b**) sequences.

Source comparisons (cluster-based permutation test) at 4 Hz indicated that the source power corresponding to T1 sequences was significantly stronger than that of the baseline condition (T1 sequences, $t_s = 4.48e+06$, $p_c < 0.004$). And the effect was mostly pronounced bilaterally in the frontal (IFG, SFG, OrG), temporal (ITG, MTG, STG, pSTS, INS) and central (PoG, PrG) regions, along with the left MFG (a larger portion of OrG at the left hemisphere was activated, see the upper panel of **Supplementary Fig. 4e**). Similarly, a significantly higher 2 Hz source power than the baseline was observed (T1 sequences, $t_s = 2.97e+06$, $p_c < 0.018$), which was largely distributed bilaterally in the frontal (MFG, OrG) and temporal (ITG, MTG, STG, INS) regions, along with the left IFG and a small portion of the right PrG (see the upper panel of **Supplementary Fig. 4d**).

More importantly, statistical estimations indicated that the source power at 1 Hz was robustly stronger for T1 sequences compared to the baseline (T1 sequences, t_s = 4.12e+06, p_c < 0.008). And the effect was largely distributed bilaterally in the frontal (MFG, OrG) and temporal (ITG, STG) areas, together with the right IFG, INS and MTG (see upper panel of **Supplementary Fig. 4c**). To assess the magnitude of the effect, paired-sample t-tests were conducted to compare the average power within the cluster at the frequencies of interest (i.e., 1, 2, and 4 Hz) with the baseline. The comparison indicated that the source power at 4 Hz (t (13) = 3.40, p < 0.004, see lower panel of **Supplementary Fig. 4e**), 2 Hz (t (13) = 4.83, p < 3.27e-04, see lower panel of **Supplementary Fig. 4d**) and 1 Hz (t (13) = 5.63, p < 8.11e-05, see lower panel of **Supplementary Fig. 4c**) were pronounced.

The power results here indicated that the brain can simultaneously track the structures at different timescales and representational level (i.e., syllables, words and learned noun pairs), which was consistent with the findings from previous studies (Ding et al., 2016; Gui et al., 2020; Jin et al., 2018; Kaufeld et al., 2020; Lu et al., 2023; Martin & Doumas, 2017; Ten Oever et al., 2022). However, it is worth noting that the top-level structures (1 Hz, noun pairs) here were not grammatical in Dutch, indicating that in this case the effect was not driven by grammatical chunking. Furthermore, we note that semantic associations could facilitate the extraction of the 1 Hz structures. However, since the noun pairs violated grammatical rules in Dutch, the semantic association between the nouns in the pair had to be formed by means of statistical cues based on experience with the stimuli. In other words, utilizing statistical cues was a prerequisite for associating the two nouns. More importantly, if semantics association was the main factor that led to the

extraction of the 1 Hz structure, the cortical origins corresponding to the 1 Hz power should have reflected it (see the upper panel of **Supplementary Fig. 4c**). Notably, the source distribution did not match with the typical patten for semantics-related processing, e.g., showing a strong left hemispheric bias (Ding et al., 2016; Giraud & Poeppel, 2012; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007). In addition, the concurrent tracking of words (2 Hz) that were defined by multiple types of cues (i.e., prosodic, linguistic and statistical) and statistically-defined noun pairs (1 Hz) revealed the flexibility of the brain in constructing representations to track during speech processing. That is to say, the power effects reflect the formation of hierarchical representations could be a manifestation of a generalized mechanism, which can be induced by any effective structural cue.

The source localizations for syllables (4 Hz, see **Supplementary Fig. 4e**) and words (2 Hz, see **Supplementary Fig. 4d**) exhibited a left hemispheric dominance, which was consistent with the spatial pattern related to speech processing (Giraud & Poeppel, 2012; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007). In contrast, the cortical origins for the noun pairs (1 Hz, see **Supplementary Fig. 4c**) were strongly biased towards the right hemisphere. Previous findings have suggested the association between the processing of statistical regularities and the right hemisphere (Corballis, 2014; Janacsek et al., 2015; Kaposvari et al., 2018; Rauch et al., 1995; Roser et al., 2011; Schapiro & Turk-Browne, 2015). The variations in hemispheric dominance orientations further suggested that the underlying mechanisms were different between building words and syllables (2 Hz and 4 Hz) and constructing noun pairs (1 Hz).

Furthermore, we noticed that the source distributions for words and syllables, when the top-level units were noun pairs (1 Hz, see **Supplementary Fig. 4d** and **Supplementary Fig. 4e**), differed from those when words were the highest-level structures (2 Hz, see **Supplementary Fig. 2c** and **Supplementary Fig. 2d**). Intuitively, the differences can be explained by the fact that more layers of units needed to be represented in the former than in the latter situation. Given the more complex hierarchical relationships, it was natural that the cortical distributions differed. However, a formal and detailed hypothesis is required to explain the underlying reasons. One potential explanation could lie in the brain's use of structured layers to process different levels of information. Topological representation is commonly accepted in neuroscience, meaning that to successfully extract information from a physical stimulus, the brain represents the required features and their compositions hierarchically. In other words, higher-level cortical regions depend on and encode combinations of lower-level representations. (Chang et al., 2010; Hickok & Poeppel, 2007; Hubel & Wiesel, 1962; Knudsen et al., 1987; Mesgarani & Chang, 2012; Mesgarani et al., 2014). In our case, therefore, it was highly likely that the required acoustic features at the bottom layer (4 Hz, syllables) for building the top-layer units varied depending on whether the highest level consisted of noun pairs (statistically defined) or words (multiple types of cues indicated the units).

However, we do not conclude that the source localizations found for syllables (4 Hz), words (2 Hz) and noun pairs (1 Hz) are the selfsame cortical networks that support the building of these representational hierarchies. Though the distributions might largely reflect the processes, there were unrelated features represented due to the topological manner of the brain. For instance, some kinds of acoustic features that reflected the

346 physical features of the speech stimulus (e.g., noise level, loudness, speaker's accent, etc.)
347 were always represented at higher layers no matter the number of hierarchies. However,
348 these features themselves were not the key gradients leading to perception of words from
349 syllables and phonemes (e.g., one cannot use noise level as a cue to a word or noun pair).
350 Therefore, the spatial distributions are not further discussed here. Instead, we will
351 attempt to isolate these unrelated factors and then address the cortical significance in a
352 later section (see the results on connectivity). This inference applies to all the univariate
353 source findings in the study.

354 Similar to previous experiments, neural phase activities were analyzed as well.
355 Statistical comparisons on T1 sequences (paired sample t-test (FDR corrected) suggested
356 that the phase coherence at 1 Hz ($t(13) = 5.96$, $p < 2.36e-05$, see the left panel of
357 **Supplementary Fig. 4f**), 2 Hz ($t(13) = 14.89$, $p < 7.51e-10$, see the left panel of
358 **Supplementary Fig. 4f**) and 4 Hz ($t(13) = 10.52$, $p < 4.93e-08$, see the left panel of
359 **Supplementary Fig. 4f**) were significantly stronger than their corresponding baselines
360 (the orange bars in **Supplementary Fig. 4f**). In contrast, a significant phase coherence
361 was only found at 4 Hz for T2 ($t(13) = 9.50$, $p < 1.62e-07$, see middle panel of
362 **Supplementary Fig. 4f**) and T3 ($t(13) = 11.51$, $p < 1.71e-08$, see the right panel of
363 **Supplementary Fig. 4f**) sequences. Once more, no statistical evidence was found to
364 suggest that the clustered phase angles were consistent across participants (see the
365 transparent lines in **Supplementary Fig. 4f**).

366 Further source reconstructions were conducted on T1 sequences for the
367 frequencies of interest. Statistical comparisons (cluster-based permutation tests) first
368 indicated that source phase synchronization was significantly stronger at 4 Hz compared

to neighboring frequency bins (T1 sequences, t_s = 7.03e+06, p_c < 0.002). And the effect was largely localized bilaterally in the frontal (IFG, MFG, SFG, OrG) and temporal (ITG, MTG, STG) regions, along with the left PrG, right INS, right IPL and right PoG (see the upper panel of **Supplementary Fig. 4i**). Similarly, a robust 2 Hz source phase coherence was identified (T1 sequences, t_s = 6.62e+06, p_c < 0.002), which was found distributed bilaterally in the frontal (IFG, MFG, SFG, OrG), temporal (ITG, MTG, STG, INS) and central (IPL, PrG, PoG, SPL) areas (see the upper panel of **Supplementary Fig. 4h**). More critically, statistical comparisons indicated a significant source phase coherence at 1 Hz (T1 sequences, t_s = 3.30e+06, p_c < 0.016). The effect was most pronounced bilaterally in the frontal (MFG, OrG) and temporal (ITG, MTG, STG, INS) regions, together with left IPL, right IFG and right PrG (see the upper panel of **Supplementary Fig. 4g)**. The temporal evolutions of the averaged phase coherence within the cluster for 1 Hz, 2 Hz and 4 Hz are shown at the lower panel of **Supplementary Fig. 4g**, **Supplementary Fig. 4h** and **Supplementary Fig. 4i**, respectively.

The strong phase coherence at 1, 2, and 4 Hz indicated that phase activities were involved in representing hierarchical structures, even when the underlying cues differed across layers and when more layers were embedded in the speech stimuli. The variations in cortical origins between neural power and phase responses suggest that different networks and/or processes may be associated with these two types of neural readouts. As emphasized previously, a generalized theoretical framework is needed (see the model) and should be validated (see results of connectivity) to further uncover the spatial significance.

Note that consistent with previous studies (Ding et al., 2016; Henin et al., 2021; Pei et al., 2023; Sheng et al., 2019), we observed a 3 Hz peak in the power spectrum, As this peak is not central to our research focus, we do not discuss it further here. However, it can be well accounted for by our theoretical model. For details, see **The 3 Hz peak in the neural response spectrum** in **the Supplementary Information.**

Neural tracking of hierarchical statistical structures.

In the previous section, we observed that hierarchical representations in the neural phase and power activity occurred when multiple layers (i.e., syllables, words and noun pairs) were involved and the cues for the triggered integration across hierarchical layers varied. To generalize this phenomenon, it is important to show that simultaneous tracking persists even when only one type of structural cue is present and consistent across different layers. To test this, two experiments were conducted. In the first one (**Experiment 5**), which was similar to its counterpart in the last section (**Experiment 3**), we trained the Dutch participants to extract 4-syllable noun pairs (1 Hz, combined by two singular nouns) in Mandarin Chinese from sequences with varying duration (1, 2 or 3 seconds, for details see **Methods**). Then, **Experiment 6** was followed, where the participants listened to the same three types of sequences (T1, T2 and T3) as in **Experiment 1**, **2** and **4**, except that the stimuli were constructed from the trained material in **Experiment 5**, with every two words forming a 1 Hz unit (without replacement). Since the participants did not understand Mandarin Chinese and the syllable sequences were isochronous, both linguistic and prosodic cues were removed. Therefore, any observed hierarchical representations should be driven by statistical

information. Consistent with the previous section, only the neural activity from **Experiment 6** was analyzed.

Statistical comparisons (paired sample t-test, FDR corrected) on the sensor-level power response of T1 sequences indicated that the induced power at 1 Hz ($t(13) = 5.68$, $p < 3.74e-05$, see **Supplementary Fig. 5b**), 2 Hz ($t(13) = 3.53$, $p < 1.84e-03$, see **Supplementary Fig. 5b**) and 4 Hz ($t(13) = 10.65$, $p < 4.31e-08$, see **Supplementary Fig. 5b**) was significantly higher than in their corresponding neighboring frequency bins. In comparison, only a 4 Hz peak occurred for T2 ($t(13) = 10.55$, $p < 4.81e-08$, see **Supplementary Fig. 5b**) and T3 ($t(13) = 7.32$, $p < 2.88e-06$, see **Supplementary Fig. 5b**) sequences.

The sensor space phase activity exhibited a similar pattern. Statistical estimations (paired sample t-test, FDR corrected) on the phase coherence corresponding to T1 sequences indicated that a stronger phase synchronization than the baseline (the orange bar in **Supplementary Fig. 5f**) occurred at 1 Hz ($t(13) = 5.93$, $p < 2.48e-05$, see the left panel of **Supplementary Fig. 5f**), 2 Hz ($t(13) = 5.08$, $p < 1.05e-04$, see the left panel of **Supplementary Fig. 5f**) and 4 Hz ($t(13) = 15.94$, $p < 3.25e-10$, see the left panel of **Supplementary Fig. 5f**). In contrast, a significant phase coherence was observed only at 4 Hz for T2 ($t(13) = 16.20$, $p < 2.66e-10$, see middle panel of **Supplementary Fig. 5f**) and T3 ($t(13) = 8.35$, $p < 6.90e-07$, see right panel of **Supplementary Fig. 5f**) sequences. And no statistical evidence was found to suggest that the phase angles were consistent across participants in any condition (see transparent lines in **Supplementary Fig. 5f**).

Source reconstructions were first applied to the power activity for T1 sequences. Statistical analysis (cluster-based permutation test) indicated that the source power at 4 Hz was significantly higher than the baseline (T1 sequences, t_s = 4.24e+06, p_c < 0.002, see the upper panel of **Supplementary Fig. 5e**). And the effect was most pronounced bilaterally in the frontal (MFG, SFG, OrG), temporal (ITG, MTG, STG, INS, pSTS) and central (PoG, PrG) areas, together with left IFG and right IPL.

Similarly, a robust 2 Hz source power was identified (T1 sequences, t_s = 2.99e+06, p_c < 0.008, see the upper panel of **Supplementary Fig. 5d**), which was bilaterally localized in the frontal (IFG, OrG) and temporal (ITG, MTG, STG, INS) regions, along with left MFG, pSTS, PrG and PoG.

More importantly, statistical comparisons indicated a significant 1 Hz source power (T1 sequences, t_s = 1.91e+06, p_c < 0.05, see the upper panel of **Supplementary Fig. 5c**), which was bilaterally distributed at STG, INS and OrG, together with left IFG, MTG, ITG and right MFG. Further checking on the magnitude of the effect (paired sample t-test) revealed that the source power at 1 Hz (t (13) = 4.97, p < 2.52e-04, see the lower panel of **Supplementary Fig. 5c**), 2 Hz (t (13) = 5.70, p < 7.23e-05, see the lower panel of **Supplementary Fig. 5d**) and 4 Hz (t (13) = 4.25, p < 9.33e-04, see the lower panel of **Supplementary Fig. 5e**) was prominent.

The last set of analyses were conducted to estimate the cortical origins of the phase synchronizations at the frequencies of interests. Statistical comparisons (cluster-based permutation test) on T1 sequences first indicated that the source-level phase coherence at 4 Hz was significantly stronger than its corresponding neighboring frequency bins (T1 sequences, t_s = 5.71e+06, p_c < 0.002, see the upper panel of **Supplementary Fig. 5i**).

460 And the effect was bilaterally localized in the frontal (MFG, OrG), temporal (ITG, MTG,
461 STG), central (PrG) and posterior (LOcC) areas, alone with the regions at left hemisphere
462 (IFG, INS, PoG) and right hemisphere (IPL, SFG, pSTS).

463 In addition, a significant 2 Hz phase coherence was detected (T1 sequences, t_s =
464 4.47e+06, p_c < 0.002, see the upper panel of **Supplementary Fig. 5h**), which was most
465 pronounced bilaterally in the temporal (ITG, MTG) and central (SPL) regions, along with
466 the areas at left (IFG, MFG, LOcC) and right hemisphere (STG).

467 Finally, statistical comparisons suggested that the source space phase coherence
468 was significantly stronger at 1 Hz compared to the baseline (T1 sequences, t_s = 4.70e+06,
469 p_c < 0.002, see the upper panel of **Supplementary Fig. 5i**). And the effect was
470 bilaterally localized in the frontal (IFG), temporal (ITG, MTG, STG, INS), central (IPL,
471 PoG) and posterior (LOcC) areas, together with the regions in left hemisphere (MFG,
472 pSTS, PrG). The averaged phase coherence (within cluster) at 1, 2 and 4 Hz is shown at
473 the lower panels of **Supplementary Fig. 5g**, **Supplementary Fig. 5h** and
474 **Supplementary Fig. 5i**, respectively.

475 The results in this section suggested that both neural phase and power activity
476 tracked the hierarchical structures (i.e., syllables, words and noun pairs) even when
477 statistical information, in the absence of comprehension, was the only accessible cue. The
478 sensor-level effects in phase and power were evoked regardless of the number of layers
479 embedded in the stimuli and despite variations in the availability of cues, again suggesting
480 that the tracking effect was not limited to being triggered by linguistic structure.

481 Although the hierarchical representations exhibited similar patterns at the sensor
482 level across different situations, the cortical origins differed, which seemed to be driven

by source of information used to track speech, i.e., the cues. In addition, the source variations across the two neural readouts (power and phase) could be associated with different networks being recruited to process specific information. The source-level findings appeared to incorporate in-depth information to discriminate different types of cues and to isolate the roles associated with the two neural measurements. In fact, similar effects at the sensor level do not necessarily project similarity in the underlying mechanisms. For instance, it is possible that noun pairs (1 Hz) were directly built upon syllables (4 Hz) when linguistic cues contradicted statistical information (e.g., **Experiment 4**), and were constructed through words (2 Hz) when they did not (e.g., **Experiment 6**). In other words, simultaneous representation of hierarchy does not imply that the underlying building processes were also progressive. Therefore, it is necessary to delve deeper into the source significance. However, as discussed previously (see the discussion in the last section), a mechanistic framework that generalizes the cortical tracking effect is needed to uncover the spatial implications (see the model section).

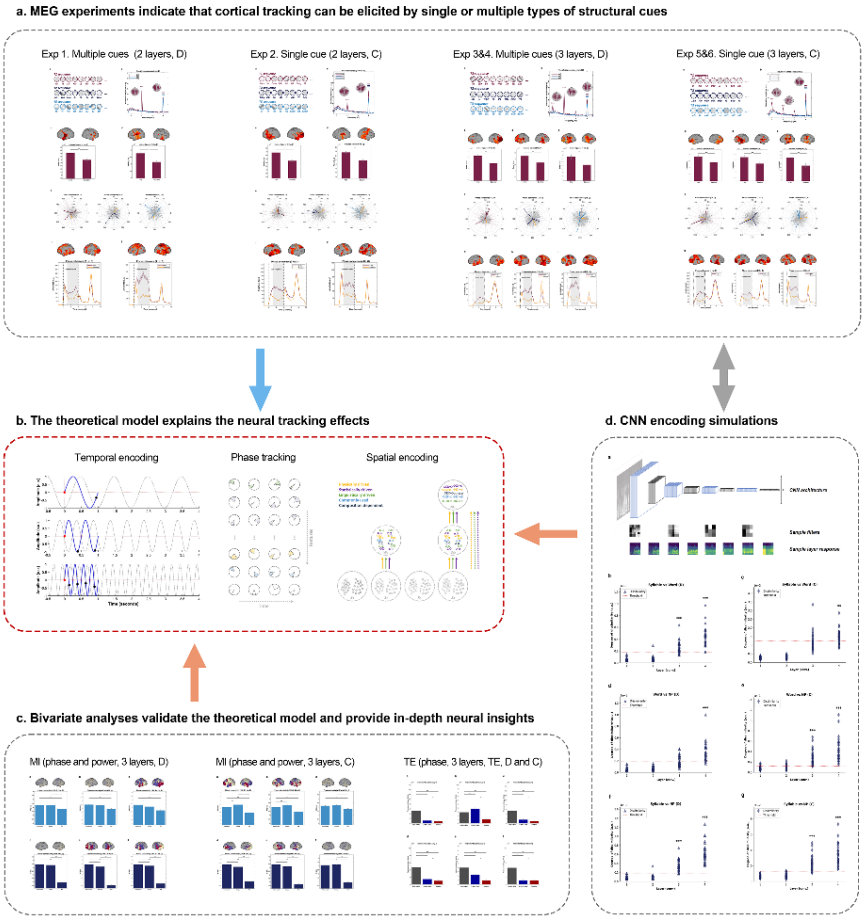
The 3 Hz peak in the neural response spectrum.

We observed a 3 Hz peak in the power response spectrum in cases where noun pairs (1 Hz) were the highest-level structures. This peak was significant compared to its neighboring bins ($p < 1.00e-03$ for all situations, see **Supplementary Fig. 4b** and **Supplementary Fig. 5b**). Since this 3 Hz rhythm did not correspond to any experimentally manipulated structures, a straightforward question arises: how did it occur, and what does the peak reflect? Previous studies exploring hierarchical representation have shown similar patterns (Ding et al., 2016; Henin et al., 2021; Pei et

al., 2023; Sheng et al., 2019), where a power response occurs at an untargeted harmonic of the fundamental frequency.

The relationship between the fundamental frequency and the 3 Hz peak did not align with the typical harmonic structure (i.e., the peak at 1 Hz was not nine times higher, 3 squared, than that at 3 Hz), suggesting that the harmonic hypothesis cannot fully explain the effect. Therefore, it is reasonable to conclude that the effect was driven by multiple factors. One plausible factor can be derived from our theoretical model. Since four syllables were presented each second (4 Hz), there were three consecutive pairs of syllables (e.g., the 1st and 2nd, 2nd and 3rd, and the 3rd and 4th) that could potentially form higher-level structures (e.g., three words). When compositionality was estimated for these syllable pairs, it followed a rhythm of three times per second (3 Hz). Thus, it is possible that the 3 Hz peak in the power spectrum was partially driven by these combinability estimations. Additionally, the 3 Hz peak might have been partly influenced by some unaccounted-for physical cues from an overall perspective. However, since this frequency was not the targeted focus and is unrelated to the study's conclusions, no further investigation was conducted on it.

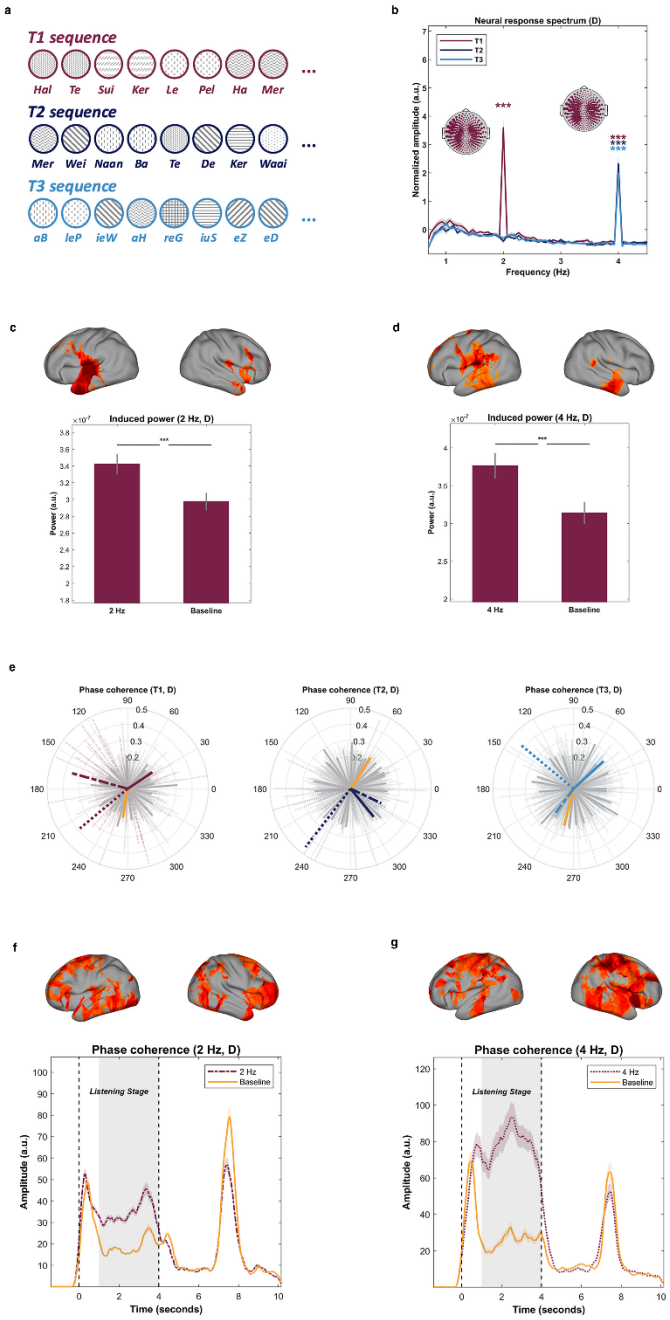
Supplementary Fig. 1. Logical framework of the study.



529 A diagram illustrating the logical framework of the study and the interrelationships among its key estimations. **a.**
530 Experimental section. *Experiment 1*. Dutch participants listened to Dutch syllable sequences, where the top-layer
531 units were words (singular nouns, 2 Hz), and multiple types of structural cues (i.e., prosodic, statistical, and
532 linguistic) were available. *Experiment 2.* Dutch participants listened to Mandarin syllable sequences, where the top-
533 layer units were also words (singular nouns, 2 Hz), but only statistical information (i.e., transitional probabilities)
534 was available to build hierarchical relationship. *Experiments 3 & 4.* Dutch participants listened to Dutch syllable
535 sequences, where the top-layer units were noun pairs (1 Hz) that violated grammatical rules. To build hierarchical
536 relationships from these sequences, multiple types of structural cues were available. *Experiments 5 & 6.* Dutch
537 participants listened to Mandarin Chinese syllable sequences, where the top-layer units were noun pairs (1 Hz), and
538 only statistical information was available to establish hierarchy. The key takeaway from these experiments is that
539 cortical tracking effects can be elicited regardless of the number of available cues (i.e., single vs. multiple) or the
540 depth of hierarchy (i.e., two vs. three levels), and the cortical distribution reflects the variation across conditions,
541 such as the availability of structural cues. **b.** Theoretical model explaining the neural observations. The theoretical
542 model was constructed based on the experimental results, isolating and integrating the roles of different neural
543 readouts (i.e., phase synchronization and power enhancement) and types of structural cues (e.g., prosodic,

statistical, and linguistic) across time, frequency, and spatial dimensions. **c.** Bivariate analyses of the neural data. Connectivity analyses were conducted on sequences (both Dutch and Mandarin) where the top-layer units were noun pairs (1 Hz). All estimations supported the theoretical model and provided in-depth evidence for how hierarchical relationships are constructed. **d.** Encoding simulations using CNNs. Based on the neural results and the theoretical model, we simulated the hierarchical encoding process using convolutional neural networks (CNNs). The simulation results validated the theoretical model, supported the neural observations, and reflected the availability of structural cue. The interconnections among various sections are depicted by arrows, where the blue, orange, and grey arrows indicate 'lead to', 'validate', and 'mutually support', respectively. D and C denote Dutch stimuli and Mandarin Chinese stimuli, respectively.

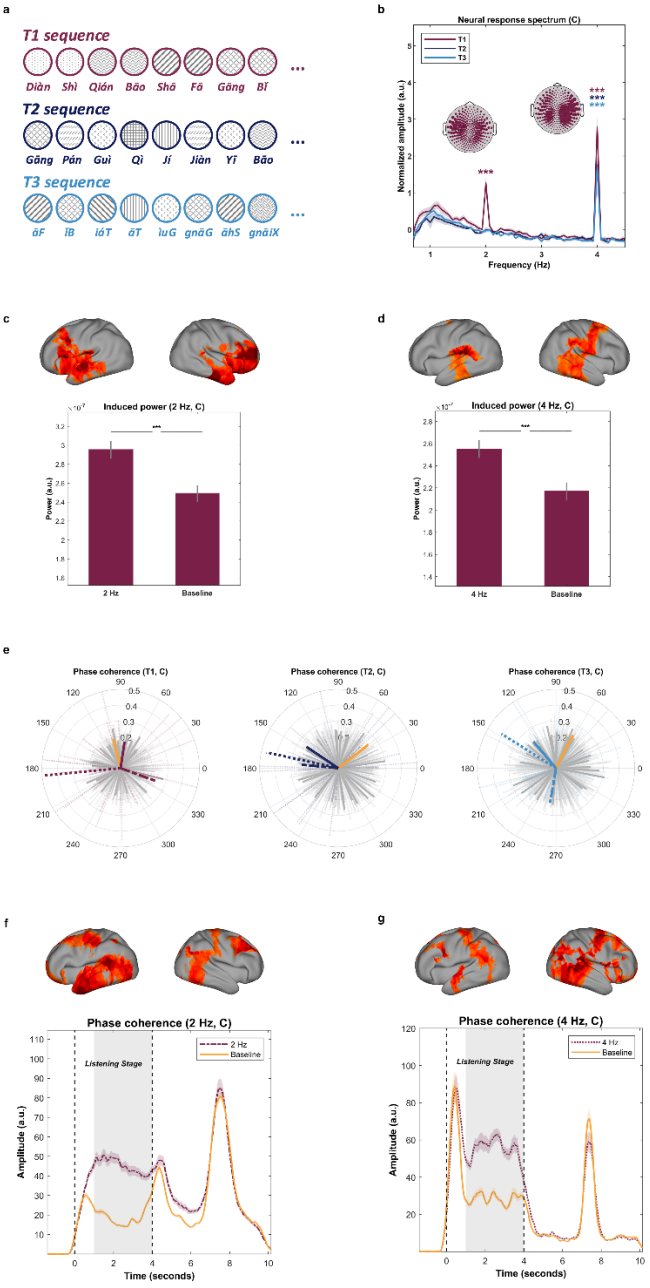
Supplementary Fig. 2. Neural tracking of linguistic structures.



a. Three types of Dutch syllable sequences were used in Experiment 1. A sample disyllabic noun sequence (T1), a random syllable sequence (T2), and a backward-played random syllable sequence (T3), are shown in red, dark blue, and light blue, respectively. In the plots, each circle represents one syllable, and the gray shading within the circles reflects the association across syllables. **b.** Neural power spectrums corresponding to the three types of sequences. A significant power peak was found at 2 Hz and 4 Hz for T1 sequences, while only a 4 Hz peak was observed for T2 and T3 sequences (three stars indicate $p < 0.005$). The shaded area on each line represents two SEMs centered around the mean. The topographies illustrate the weights of each sensor when spatially extracting the optimized

neural response at 2 Hz and 4 Hz (for T1 sequences only). **c, d.** The cortical surface plots display the source power localizations, with the left and right plots representing the left and right hemispheres, respectively. The red areas mark the regions with pronounced activity, where darker colors indicate higher t-values. The lower panels show the magnitude of the source power effect, with the gray bars depicting 2 SEMs centered around the means. **e.** The left, middle, and right panels show the sensor space phase coherence for T1, T2, and T3 sequences, respectively. The averaged phase coherence at 1, 2, and 4 Hz is shown with solid, dash-dotted, and dotted lines, respectively. Significant phase coherence was identified at 2 Hz and 4 Hz for T1 sequences, while only a significant 4 Hz phase coherence was detected for T2 and T3 sequences. The statistical baseline is indicated by the orange lines. The averaged phase coherence at all other frequencies is shown in solid gray lines, and the individual-level phase coherence is depicted by transparent thinner lines. **f, g.** The brain surface plots depict the cortical localization of the averaged phase coherence (1 to 4 seconds after audio onset). The red areas highlight the detected pronounced regions, with darker red colors indicating higher t-values. The lower panels show the averaged phase coherence (within the cluster) over time. The shaded area in each line represents 2 SEMs centered around the mean.

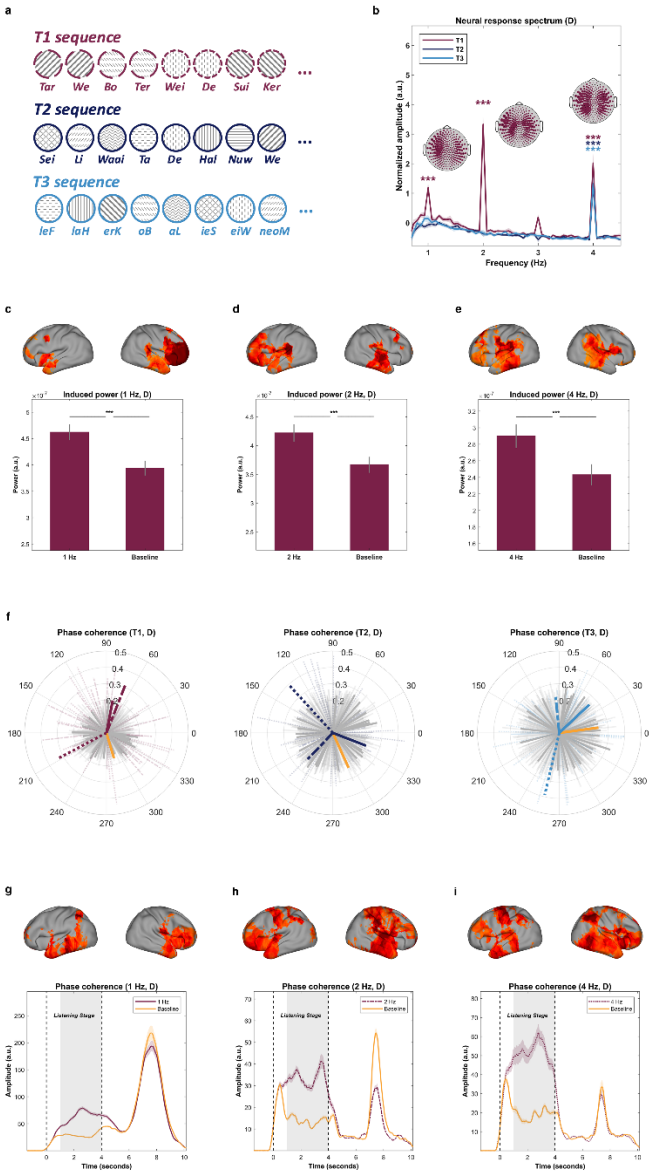
Supplementary Fig. 3. Neural tracking of statistical structures.



a. Three types of Mandarin Chinese syllable sequences were used in Experiment 2. A sample disyllabic noun sequence (T1), a random syllable sequence (T2), and a backward-played random syllable sequence (T3), are shown in red, dark blue, and light blue, respectively. In the plots, each circle represents one syllable, and the gray shading within the circles reflects the association across syllables. **b.** Neural power spectrums corresponding to the three types of sequences. A significant power peak was found at 2 Hz and 4 Hz for T1 sequences, while only a 4 Hz peak was observed for T2 and T3 sequences (three stars indicate $p < 0.005$). The shaded area on each line represents two SEMs centered around the mean. The topographies illustrate the weights of each sensor when spatially extracting the optimized neural response at 2 Hz and 4 Hz (for T1 sequences only). **c, d.** The cortical surface plots

display the source power localizations, with the left and right plots representing the left and right hemispheres, respectively. The red areas mark the regions with pronounced activity, where darker colors indicate higher t-values. The lower panels show the magnitude of the source power effect, with the gray bars depicting 2 SEMs centered around the means. **e.** The left, middle, and right panels show the sensor space phase coherence for T1, T2, and T3 sequences, respectively. The averaged phase coherence at 1, 2, and 4 Hz is shown with solid, dash-dotted, and dotted lines, respectively. Significant phase coherence was identified at 2 Hz and 4 Hz for T1 sequences, while only a significant 4 Hz phase coherence was detected for T2 and T3 sequences. The statistical baseline is indicated by the orange lines. The averaged phase coherence at all other frequencies is shown in solid gray lines, and the individual-level phase coherence is depicted by transparent thinner lines. **f, g.** The brain surface plots depict the cortical localization of the averaged phase coherence (1 to 4 seconds after audio onset). The red areas highlight the detected pronounced regions, with darker red colors indicating higher t-values. The lower panels show the averaged phase coherence (within the cluster) over time. The shaded area in each line represents 2 SEMs centered around the mean.

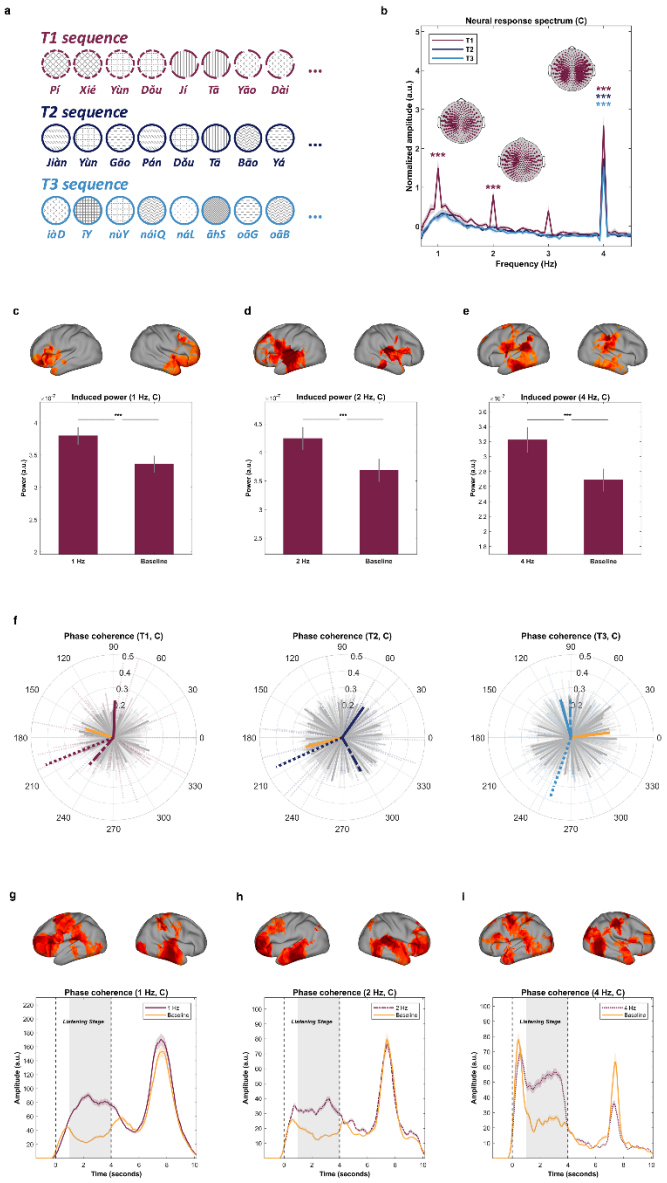
Supplementary Fig. 4. Neural tracking of linguistic and statistical structures.



a. Three types of Dutch syllable sequences were used in Experiment 4. A sample disyllabic noun sequence (T1), a random syllable sequence (T2), and a backward-played random syllable sequence (T3), are shown in red, dark blue, and light blue, respectively. In the plots, each circle represents one syllable, and the gray shading within the circles reflects the association across syllables. Due to the training in Experiment 3, every two consecutive disyllabic nouns (or four syllables; without replacement) in each stimulus could be statistically combined to form a noun pair (1 Hz units). The association at 1 Hz across syllables is indicated by the shape of the outlines of the circles. **b.** Neural power spectrums corresponding to the three types of sequences. A significant power peak was found at 1, 2, and 4 Hz for T1 sequences, while only a 4 Hz peak was observed for T2 and T3 sequences (three stars indicate $p < 0.005$). The shaded area on each line represents two SEMs centered around the mean. The topographies illustrate the weights of each sensor in spatially extracting the optimized neural response at 1, 2, and 4 Hz (for T1 sequences

only). **c, d, e.** The cortical surface plots display the source power localizations, with the left and right plots representing the left and right hemispheres, respectively. The red areas mark the regions with pronounced activity, where darker colors indicate higher t-values. The lower panels show the magnitude of the source power effect, with the gray bars depicting 2 SEMs centered around the means. **f.** The left, middle, and right panels show the sensor space phase coherence for T1, T2, and T3 sequences, respectively. The averaged phase coherence at 1, 2, and 4 Hz is shown with solid, dash-dotted, and dotted lines, respectively. Significant phase coherence was detected at 1, 2, and 4 Hz for T1 sequences, whereas only significant 4 Hz phase coherence was observed for T2 and T3 sequences. The statistical baseline is indicated by the orange lines. The averaged phase coherence across participants at all other frequencies is shown in solid gray lines, while the individual-level phase coherence is depicted by transparent thinner lines. **g, h, i.** The brain surface plots depict the cortical localization of the averaged phase coherence (1 to 4 seconds after audio onset). The red areas highlight the pronounced regions (with darker red colors indicating higher t-values). The lower panels show the averaged phase coherence (within the cluster) over time. The shaded area in each line covers 2 SEMs centered around the mean.

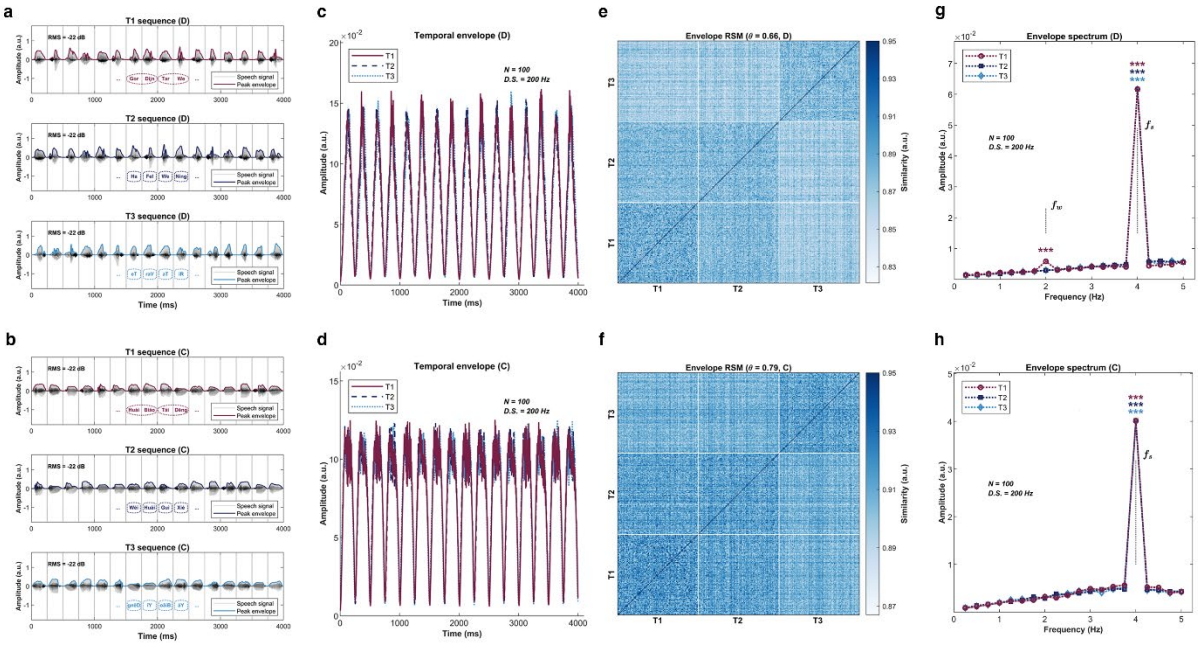
Supplementary Fig. 5. Neural tracking of hierarchical statistical structures.



a. Three types of Mandarin Chinese syllable sequences were used in Experiment 6. A sample disyllabic noun sequence (T1), a random syllable sequence (T2), and a backward-played random syllable sequence (T3), are shown in red, dark blue, and light blue, respectively. In the plots, each circle represents one syllable, and the gray shading within the circles reflects the association across syllables. Due to the training in Experiment 5, every two consecutive disyllabic nouns (or four syllables, without replacement) in each sequence could be statistically combined to form a novel compound (1 Hz units). The association at 1 Hz across syllables is indicated by the shape of the outlines of the circles. **b.** Neural power spectrums corresponding to the three types of sequences. A significant power peak was found at 1, 2, and 4 Hz for T1 sequences, while only a 4 Hz peak was observed for T2 and T3 sequences (three stars indicate $p < 0.005$). The shaded area on each line represents two SEMs centered around the mean. The topographies illustrate the weights of each sensor in spatially extracting the optimized

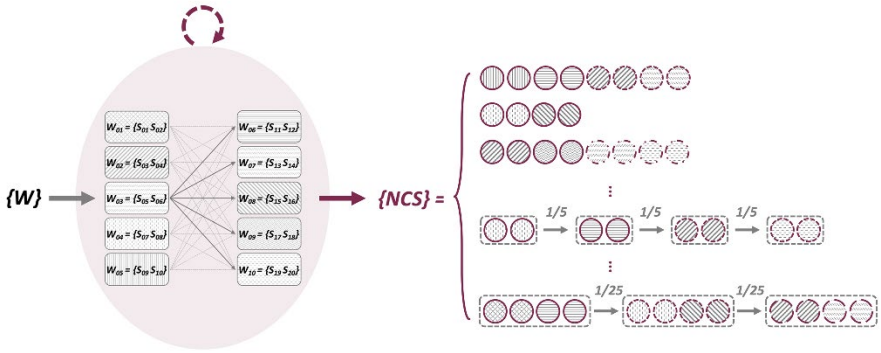
neural response at 1, 2, and 4 Hz (for T1 sequences only). **c, d, e.** The cortical surface plots display the source power localizations, with the left and right plots representing the left and right hemispheres, respectively. The red areas mark the regions with pronounced activity, where darker colors indicate higher t-values. The lower panels show the magnitude of the source power effect, with the gray bars depicting 2 SEMs centered around the means. **f.** The left, middle, and right panels show the sensor space phase coherence for T1, T2, and T3 sequences, respectively. The averaged phase coherence at 1, 2, and 4 Hz is shown with solid, dash-dotted, and dotted lines, respectively. Significant phase coherence was detected at 1, 2, and 4 Hz for T1 sequences, whereas only significant 4 Hz phase coherence was observed for T2 and T3 sequences. The statistical baseline is indicated by the orange lines. The averaged phase coherence across participants at all other frequencies is shown in solid gray lines, while the individual-level phase coherence is depicted by transparent thinner lines. **g, h, i.** The brain surface plots depict the cortical localization of the averaged phase coherence (1 to 4 seconds after audio onset). The red areas highlight the pronounced regions (with darker red colors indicating higher t-values). The lower panels show the averaged phase coherence (within the cluster) over time. The shaded area in each line covers 2 SEMs centered around the mean.

Supplementary Fig. 6. Acoustic normalization and analysis.



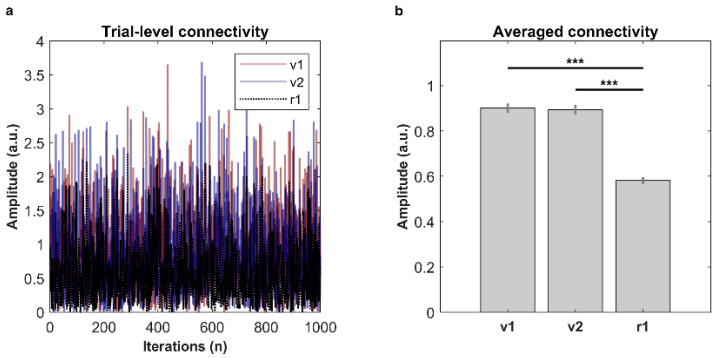
a. Temporal dynamics of Dutch sample sequences used in the study. The upper, middle, and lower panels show the T1, T2, and T3 sequences, respectively. In the figure, the black transparent lines represent the actual waveforms of the speech stimuli over time, and the colored solid lines indicate the corresponding temporal envelopes. To optimize the isochronicity of the stimuli, each syllable was truncated or zero-padded to 0.25 seconds and then tapered at both ends (5%), and its RMS value was normalized to -22 dB. The duration of each syllable is marked by black dotted-line enclosed rectangles. **c.** The red, dark blue, and light blue lines show the averaged temporal envelopes (100 sequences, resampled to 200 Hz) of T1, T2, and T3 sequences, respectively. The colored shading around each line represents 2 SEMs centered around the mean. **d.** To show the physical match across different types of sequences over time, a representational similarity matrix (RSM) was estimated. Specifically, the cosine similarity between each possible pair of sequences was calculated and tested against a permutation-derived threshold (θ = 0.66). Statistical analysis indicated that the sequences were physically well-matched both within and across types in the time domain. **g.** The red, dark blue, and light blue lines show the averaged power spectrum corresponding to T1, T2, and T3 sequences, respectively. The colored shading on each line represents 2 SEMs centered around the mean. Statistical comparisons indicated that the power at 2 Hz and 4 Hz was significantly stronger than their neighboring frequency bins for T1 sequences (three stars indicate $p < 0.005$). In contrast, only a significant 4 Hz power was identified for T2 and T3 sequences. **c, d, f, h.** The same normalizations and analyses were applied to the Mandarin Chinese stimuli. The comparisons corresponding to the Dutch counterparts are shown in the lower panel. Similar results were obtained for the Mandarin Chinese stimuli, except that only significant 4 Hz power was detected in the temporal envelopes of the three types of sequences.

Supplementary Fig. 7. Recurrent probabilistic model for generating sequences in training.



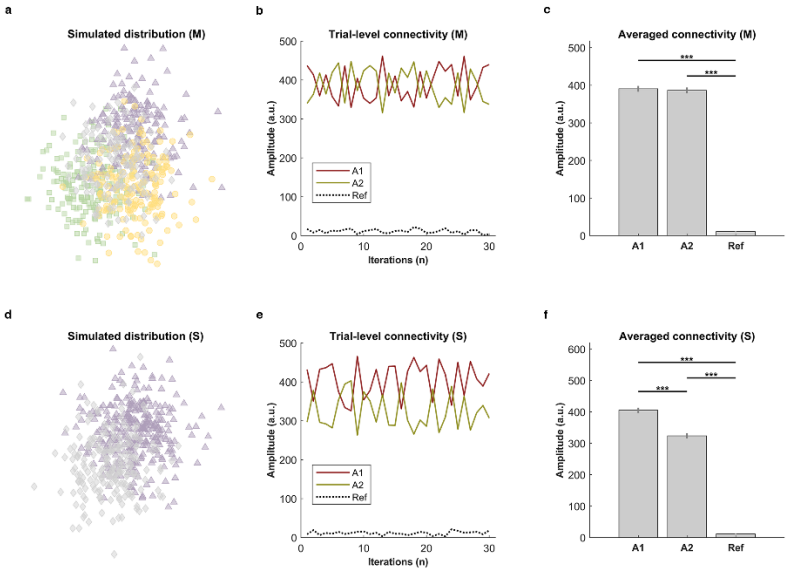
For each individual, a set of ten disyllabic words was randomly sampled from a pool of twenty. These ten words were then stochastically arranged into two sets of five. The full combination of these sets (5 × 5) yielded twenty-five 4-syllable noun-pair structures (1 second each). To embed statistical cues at multiple hierarchical levels (e.g., word and noun-pair layers) for learning, we employed a Markovian probabilistic framework to generate syllable sequences, each containing one, two, or three noun pairs. This approach ensured that the transitional probability (TP) between words within a noun pair was 1/5, and the TP between noun pairs (above the word level) was 1/25, from an overall perspective. In the figure, {w} represents the selected set of ten words, the red transparent oval illustrates the recurrent probabilistic model, and {NCS} denotes the set of generated sequences used for training. The rightmost panel displays several sample sequences, with statistical associations between words and between noun pairs indicated by grey arrows. Grey shading within the circles marks grouping at the word level (2 Hz; two syllables per word), while the shapes of the circles' outlines indicate syllable associations at the noun-pair level (1 Hz; four syllables per noun pair).

Supplementary Fig. 8. Simulations of connectivity across layers.



To demonstrate how strong connectivity occurred between layers when the response at one unit in a higher layer was a linear combination of the units layered below, simulations were conducted. Specifically, we generated 1,000 pairs of vectors (n = 10) and computed the corresponding difference vectors for each pair. Connectivity was estimated by calculating the cosine similarity between the resulting difference vectors and their components. A statistical reference was established using the same connectivity measurement between the difference vectors and randomly generated vectors (on the same scale). **a.** Trial-level simulation results. The red, blue, and black dotted lines show the connectivity between the difference vectors and their first component (v1), second component (v2), and random vectors (r1), respectively. A clear pattern can be observed, where the degree of connectivity in the targeted conditions (v1 and v2) is higher than in the reference condition (r1). **b.** To make statistical inferences, a one-way ANOVA was conducted to compare the means of the three conditions. The analysis indicated a significant difference across the means ($F(2, 2997) = 95.06$, $p < 9.37e\text{-}41$). Further pairwise comparisons using paired-sample t-tests (FDR corrected) showed that the degree of connectivity for v1 ($t(999) = 14.60$, $p < 6.53e\text{-}44$) and v2 ($t(999) = 13.98$, $p < 1.03e\text{-}40$) was significantly stronger than for r1, and the connectivity for v1 and v2 was comparable ($t(999) = 0.19$, $p < 0.84$). In the figure, the gray bars represent 2 SEMs centered around the means.

Supplementary Fig. 9. Simulations of cross-layer connectivity in space.



To illustrate the spatial mechanism for cross-layer communication and validate the neural connectivity results, simulations were conducted under conditions where one or multiple types of structural cues were accessible. **a.** The spatial distribution of the integrations of acoustic features (mimicking cortical distribution) tracked by various types of cues. In the figure, green, purple, yellow, and grey symbols represent associations driven by linguistic, statistical, physical, and composition-dependent information, respectively. The spatial relationships among these various types of representations were simulated using 2D Gaussian, where the means along the horizontal and vertical axes were set to one for cue-based representations (i.e., linguistic, statistical, and physical information) and zero for dependent ones (e.g., noise). The standard deviations were normalized to one for both axes and remained consistent across all types. The ratio between cue based (e.g., linguistic, statistical, and physical information) and dependent representations was set at 1.5 (6/4) to simulate the real hierarchical building process, reflecting that the number of structural representations exceeded that of dependent features' combinations. **b.** Trial-level simulation results. To test whether a dominant distribution would emerge when multiple structural cues were accessible, a randomized simulation was conducted 30 times. In each trial, we randomly split the spatial distribution into two parts (see Supplementary Fig. a), where one part contained 40% to 60% (50% on average) of all representations (symbols). Since each symbol represents an integration of tracked acoustic features, its activation was modeled as a weighted combination of its components. According to our model, we assigned a random weight between 0.9 and 1 for cue-based representations and between 0 and 0.2 for dependent ones. We then calculated the average connectivity (cosine similarity) of all symbols to their components in the first (A1) and second part (A2), and compared these to a statistical reference, computed as the connectivity of all symbols with randomly generated vectors (Ref). In Supplementary Fig. b, the red, yellow, and black-dotted lines represent the simulated connectivity for A1, A2, and the reference, respectively. **c.** Averaged connectivity across all conditions. A one-way ANOVA revealed a significant difference across the three conditions ($F(2, 87) = 1462.05$, $p < 1.10e\text{-}67$). Further pairwise comparisons (paired sample t-test, FDR corrected) indicated that the degree of connectivity for A1 ($t(29) = 55.56$, $p < 5.19e\text{-}31$) and A2 ($t(29) = 52.27$, $p < 3.00e\text{-}30$) was significantly stronger than for the reference, while connectivity between A1 and A2 was comparable ($t(29) = 0.74$, $p = 0.46$). **d, e, f.** The figures present the

counterpart results for simulations where only one type of structural cue was available. The parameters (e.g., the ratio, number of randomizations, and method for calculating connectivity) were identical to those used in the multi-cue simulation. Statistical analysis revealed a significant difference across the means ($F(2, 87) = 1546.46$, $p < 1.03e\text{-}68$), and pairwise comparisons indicated that the connectivity for A1 ($t(29) = 60.54$, $p < 4.40e\text{-}32$) and A2 ($t(29) = 48.68$, $p < 2.32e\text{-}29$) was significantly stronger than the reference. More importantly, the connectivity strength for A1 was significantly higher than that for A2 ($t(29) = 5.76$, $p < 3.08e\text{-}06$).

Supplementary Table 1. Dutch materials.

tij	ger	tijger	tiger
ta	fel	tafel	table
la	waai	lawaai	noise
var	ken	varken	pig
be	zem	bezem	broom
tar	we	tarwe	wheat
hal	te	halte	station
ba	naan	banaan	banana
ri	vier	rivier	river
wei	de	weide	pasture
gor	dijn	gordijn	curtain
ze	nuw	zenuw	nerve
sei	zoen	seizoen	season
sui	ker	suiker	sugar
bo	ter	boter	butter
li	moen	limoen	lemon
ko	ning	koning	king
ha	mer	hamer	hammer
le	pel	lepel	spoon
wor	tel	wortel	carrot

Supplementary Table 2. Mandarin Chinese materials

怀 (huái)	表 (biǎo)	怀表 (huái biǎo)	pocket watch
键 (jiàn)	盘 (pán)	键盘 (jiàn pán)	keyboard
相 (xiàng)	机 (jī)	相机 (xiàng jī)	camera
电 (diàn)	视 (shì)	电视 (diàn shì)	televison
熨 (yùn)	斗 (dǒu)	熨斗 (yùn dǒu)	iron
衣 (yī)	柜 (guì)	衣柜 (yī guì)	wardrobe
冰 (bīng)	箱 (xiāng)	冰箱 (bīng xiāng)	refrigerator
吉 (jí)	他 (tā)	吉他 (jí tā)	guitar
沙 (shā)	发 (fā)	沙发 (shā fā)	sofa
帐 (zhàng)	篷 (péng)	帐篷 (zhàng péng)	tent
腰 (yāo)	带 (dài)	腰带 (yāo dài)	belt
牙 (yá)	膏 (gāo)	牙膏 (yá gāo)	toothpaste
钢 (gāng)	笔 (bǐ)	钢笔 (gāng bǐ)	pen
篮 (lán)	球 (qiú)	篮球 (lán qiú)	basketball
汽 (qì)	车 (chē)	汽车 (qì chē)	car
围 (wéi)	巾 (jīn)	围巾 (wéi jīn)	scarf
台 (tái)	灯 (dēng)	台灯 (tái dēng)	table lamp
钱 (qián)	包 (bāo)	钱包 (qián bāo)	wallet
耳 (ěr)	环 (huán)	耳环 (ěr huán)	earring
皮 (pí)	鞋 (xié)	皮鞋 (pí xié)	leather shoe

References.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, *13*(11), 1428-1432.

Corballis, M. C. (2014). Left brain, right brain: facts and fantasies. *PLoS Biology*, *12*(1), e1001767.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, *19*(1), 158-164.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, *15*(4), 511-517.

Gui, P., Jiang, Y., Zang, D., Qi, Z., Tan, J., Tanigawa, H., Jiang, J., Wen, Y., Xu, L., & Zhao, J. (2020). Assessing the depth of language processing in patients with disorders of consciousness. *Nature neuroscience*, *23*(6), 761-770.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual review of neuroscience*, *37*, 347-362.

Henin, S., Turk-Browne, N. B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., & Melloni, L. (2021). Learning hierarchical sequence representations across human cortex and hippocampus. *Science Advances*, *7*(8), eabc4530.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393-402.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Janacsek, K., Ambrus, G. G., Paulus, W., Antal, A., & Nemeth, D. (2015). Right hemisphere advantage in statistical learning: evidence from a probabilistic sequence learning task. *Brain stimulation*, *8*(2), 277-282.

Jin, P., Zou, J., Zhou, T., & Ding, N. (2018). Eye activity tracks task-relevant structures during speech and auditory sequence perception. *Nature communications*, *9*(1), 5374.

Kaposvari, P., Kumar, S., & Vogels, R. (2018). Statistical learning signals in macaque inferior temporal cortex. *Cerebral cortex*, *28*(1), 250-266.

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *Journal of Neuroscience*, *40*(49), 9467-9475.

Knudsen, E. I., Lac, S. d., & Esterly, S. D. (1987). Computational maps in the brain. *Annual review of neuroscience*, *10*(1), 41-65.

Lu, Y., Jin, P., Ding, N., & Tian, X. (2023). Delta-band neural tracking primarily reflects rule-based chunking instead of semantic relatedness between words. *Cerebral cortex*, *33*(8), 4448-4458.

Martin, A. E., & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology*, *15*(3), e2000663.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233-236.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006-1010.

Pei, C., Qiu, Y., Li, F., Huang, X., Si, Y., Li, Y., Zhang, X., Chen, C., Liu, Q., & Cao, Z. (2023). The different brain areas occupied for integrating information of hierarchical linguistic units: a study based on EEG and TMS. *Cerebral cortex*, *33*(8), 4740-4751.

Rauch, S. L., Savage, C. R., Brown, H. D., Curran, T., Alpert, N. M., Kendrick, A., Fischman, A. J., & Kosslyn, S. M. (1995). A PET investigation of implicit and explicit sequence learning. *Human brain mapping*, *3*(4), 271-286.

Roser, M. E., Fiser, J., Aslin, R. N., & Gazzaniga, M. S. (2011). Right hemisphere dominance in visual statistical learning. *Journal of cognitive neuroscience*, *23*(5), 1088-1099.

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*(1), 27-52.

Schapiro, A., & Turk-Browne, N. (2015). Statistical learning. *Brain mapping*, *3*, 501-506.

Sheng, J., Zheng, L., Lyu, B., Cen, Z., Qin, L., Tan, L. H., Huang, M.-X., Ding, N., & Gao, J.-H. (2019). The cortical maps of hierarchical linguistic structures during speech perception. *Cerebral cortex*, *29*(8), 3232-3240.

Ten Oever, S., Carta, S., Kaufeld, G., & Martin, A. E. (2022). Neural tracking of phrases in spoken language comprehension is automatic and task-dependent. *Elife*, *11*, e77468.