Hierarchical dynamic coding coordinates speech comprehension in the brain.

Laura Gwilliams[1, 2], Alec Marantz[2, 5] David Poeppel[2, 6] & Jean-Remi King[3, 4]

[1]Department of Psychology, Stanford University

[2]Department of Psychology, New York University,

[3]Ecole Normale Superieure, PSL, CNRS,

[4] Meta AI,

[5]Department of Linguistics, New York University,

[6]Ernst Strungman Institute

Abstract

Speech comprehension requires the human brain to transform an acoustic waveform into meaning. To do so, the brain generates a hierarchy of features that converts the sensory input into increasingly abstract language properties. However, little is known about how these hierarchical features are generated and continuously coordinated. Here, we propose that each linguistic feature is dynamically represented in the brain to simultaneously represent successive events. To test this 'Hierarchical Dynamic Coding' (HDC) hypothesis, we use time-resolved decoding of brain activity to track the construction, maintenance, and integration of a comprehensive hierarchy of language features spanning acoustic, phonetic, sub-lexical, lexical, syntactic and semantic representations. For this, we recorded 21 participants with magnetoencephalography (MEG), while they listened to two hours of short stories. Our analyses reveal three main findings. First, the brain incrementally represents and simultaneously maintains successive features. Second, the duration of these representations depend on their level in the language hierarchy. Third, each representation is maintained by a dynamic neural code, which evolves at a speed commensurate with its corresponding linguistic level. This HDC preserves the maintenance of information over time while limiting the interference between successive features. Overall, HDC reveals how the human brain continuously builds and maintains a language hierarchy during natural speech comprehension, thereby anchoring linguistic theories to their biological implementations.

Keywords: speech, hierarchy, timescales, brain, language, machine learning, decoding

Introduction

To achieve speech comprehension, the brain generates a hierarchy of features that span a representational hierarchy from sound to meaning [1,2]. The resulting phonetic [3,4], syllabic [5,6] and lexical features [7–9] as well as their associated syntactic and compositional structures [10–13] are represented in the temporal, parietal and prefrontal cortices, with more abstract linguistic representations being encoded in more distributed and higher-level activation patterns [14,15].

However, the neural *dynamics* of this hierarchical processing remain elusive [16]. This gap of knowledge is partly due to technical limitations: functional Magnetic Resonance Imaging (fMRI) has limited temporal resolution and intracranial recordings have generally limited coverage. Consequently, how linguistic features are continuously built and maintained in this cortical hierarchy remains largely unknown [17–19].

In particular, speech comprehension hinges on a delicate balance between two opposing constraints. The first constraint involves *maintaining* low-level elements – like phonemes, long enough to integrate them into more complex units – like words [7]. The second constraint requires continuously *updating* each element to keep up with the continuous unfolding of speech, and appropriately process new incoming information. These constraints theoretically apply across all levels of the hierarchy: from assembling phonemes into words, to assembling words into sentences. A computational framework is thus essential to explain how the cortex simultaneously *maintains* and *updates* each of the representations of language to integrate increasingly high-level representations.

Time-resolved decoding of brain activity may provide a promising tool to resolve this issue [20–22]. By decoding the representations at each point in time, acoustic-phonetic [4] and visual features [23] have recently been shown to be embedded in a dynamic neural code. Specifically, temporal generalization analyses show that a representation may be maintained over long time periods by successively recruiting distinct neural populations (Figure 1). We propose that this dynamic coding can be applied hierarchically to both maintain and update the many representations of language, while avoiding interference across successive phonemes, syllables, and words; henceforth referred to as Hierarchical Dynamic Coding (HDC).

To put the HDC hypothesis to the test, we recorded magneto-encephalography (MEG) from 21 participants listening to two hours of audio stories. We fit linear models [20] to decode a comprehensive set of 54 linguistic features organized into six levels of representation: phonetic, sub-lexical, lexical, syntactic operation, lexical semantic and syntactic state. We address three main questions: (i) can we decode the features of the language hierarchy? (ii) what are their relative onsets and durations? and (iii) does their underlying neural code evolve over time, with speed commensurate to their level in the hierarchy (Figure 1)?

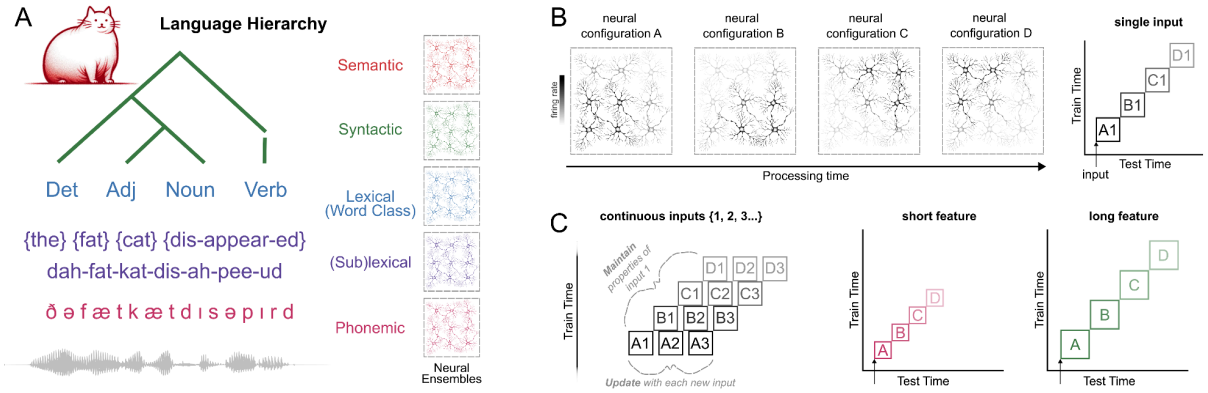

Figure 1. The Hierarchical Dynamic Coding Hypothesis. A: Schematic of the language hierarchy, from the acoustic input to the meaning of the utterance. Each feature is hypothesized to be encoded by a distinct neural ensemble. B: Schematic of the HDC

hypothesis: for each feature of the hierarchy, encoding evolves across different neural ensembles as a function of time. C: Schematic decoding result for a single speech input. Each neural configuration {A, B, C, D} is engaged in sequence, leading to a lack of generalization across all train/test times. D: Schematic decoding result for sequences of inputs, which satisfies both the constraint to maintain information over time, and update as new inputs are received. E: Schematic prediction that shorter and longer features of language will display distinct processing dynamics.

Results

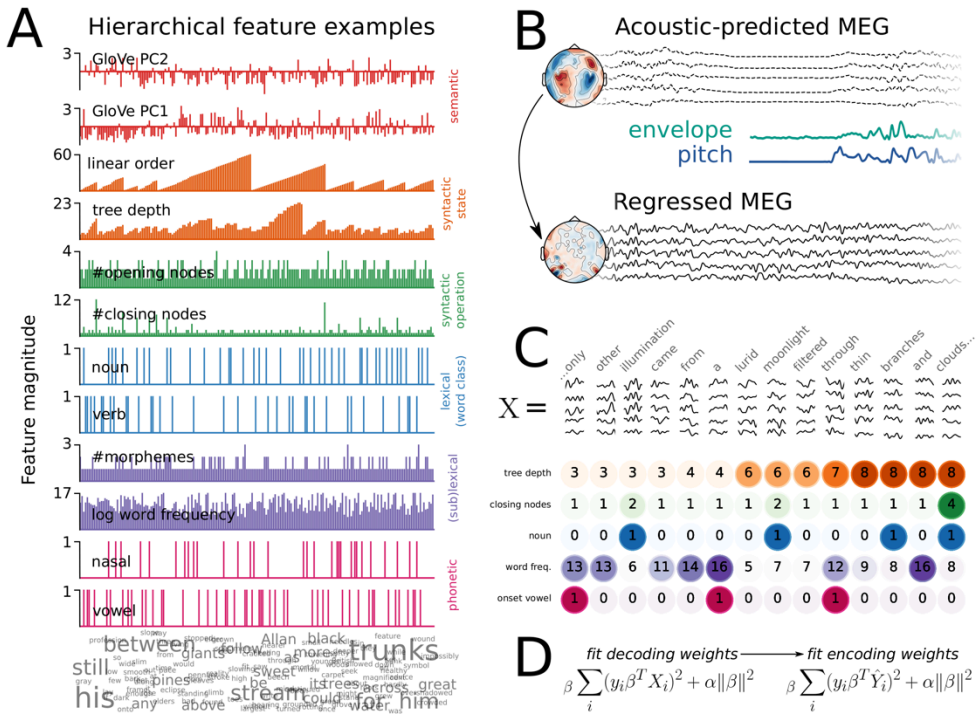

Figure 2: Methods. A: Feature values are plotted for a selection of 500 consecutive words, selecting two example features for each level (from a total of 54 features). Colour corresponds to the language level it operationalises. Word cloud at the bottom displays each of the 500 words with fontsize proportional to the number of instances in the story snippet. B: First we fit a receptive field model based on the envelope and pitch of the acoustic speech input, and regress this out of the continuous MEG signal. Topography above shows sensor weights sensitive to acoustic features; topography below shows that sensor weights are at zero after this procedure. C: Data structure. Epoch data matrix *X* has shape words (8,000) × sensors (208) × time (201). Here a schematic of the epochs is shown for a subset of the story. Below, a sample of 5 example features are displayed for each word. The superimposed number and color intensity corresponds to the feature value at a particular word. D: Main equations for the back-to-back regression method. A = the fitted regularization parameter. B = the fitted model coefficients of interest. First we fit ridge regression to decode each feature *y* from the MEG signal *X*, then we evaluate the prediction of the features ($\hat{Y}$) against the true value of *y* with an encoding model.

1.1 Robust decoding of speech features

Our first question is whether the rich suite of linguistic features can be decoded from MEG activity during continuous listening. To evaluate this, we compute the time course of each linguistic feature (Figure 2) and evaluate statistical significance with a temporal permutation cluster test of the distribution of beta coefficients across participants. Overall, our results show that we can precisely track a remarkably diverse – although not exhaustive – set of linguistic features from MEG activity (Figure 3). The results of the full statistical analyses on all features are provided in Supplementary Tables 1-12. For the analysis on raw Spearman R correlation rather than B2B regression, see Supplement.

To summarize the levels of the hierarchy that are robustly encoded in neural activity, we grouped decoding performance of the original 54 linguistic features into 6 feature families, and took the average of the features in a family, to plot 6 decoding time-courses (Figure 4A). Features across all six feature families could be detected from MEG responses, with notable differences in latency and duration: on

average, phonetic features were detectable from -40:230 ms ($t^{\wedge}$ (average t-value in the cluster) = 2.57, p = .013) relative to word offset; sub-lexical features from – 130:550 ms ($t^{\wedge}$ = 2.6, p = .002); lexical word class from -170:200 ms ($t^{\wedge}$ = 2.18, p = .029); syntactic operation from -190:1200 ms ($t^{\wedge}$ = 2.7, p < .001); syntactic state ($t^{\wedge}$ = 3.54, p < .001) and semantic vectors ($t^{\wedge}$ = 3.46, p < .001) throughout the entire search window (Figure 4A). These results are consistent across the two recording sessions (Supplementary Figure 6), thus demonstrating internal replicability (see Supplement for detailed results). Overall, this analysis confirms that, during continuous speech listening, the brain builds a rich set of hierarchical linguistically motivated features.

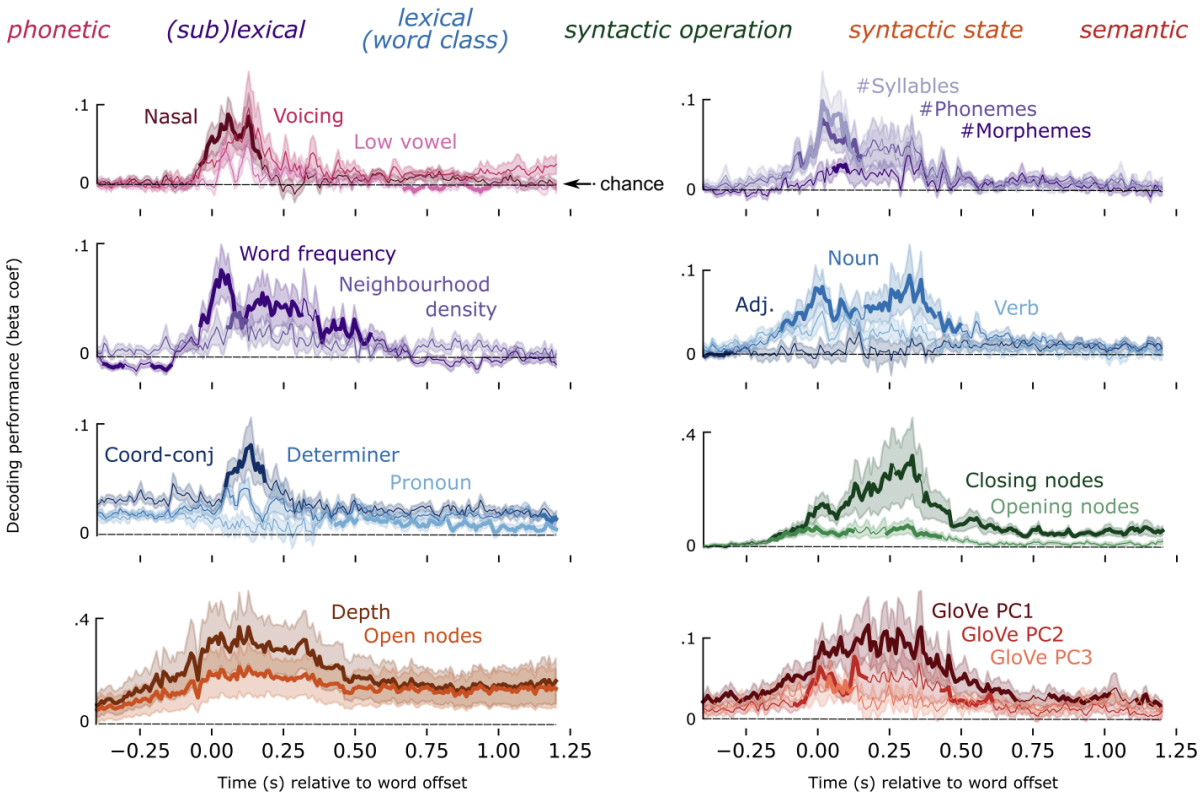

Figure 3: Feature decoding timecourses. Timecourses of decoding performance for a subset of language properties, locked to word offset. Line color corresponds to family assignment, which are listed above. Solid trace corresponds to mean performance across subjects; shading is the standard error of the mean across subjects. A bold mean trace corresponds to the result of a temporal permutation cluster test, indicating when the feature is decodable significantly better than chance. Dashed black line corresponds to chance-level performance.

1.2. The timing of linguistic representations depends on their level in the language hierarchy.

How do the latency and duration of each feature relate to their respective level in the linguistic hierarchy? To address this issue, we analyzed the average time-course of each of the 6 feature families (Figure 4).

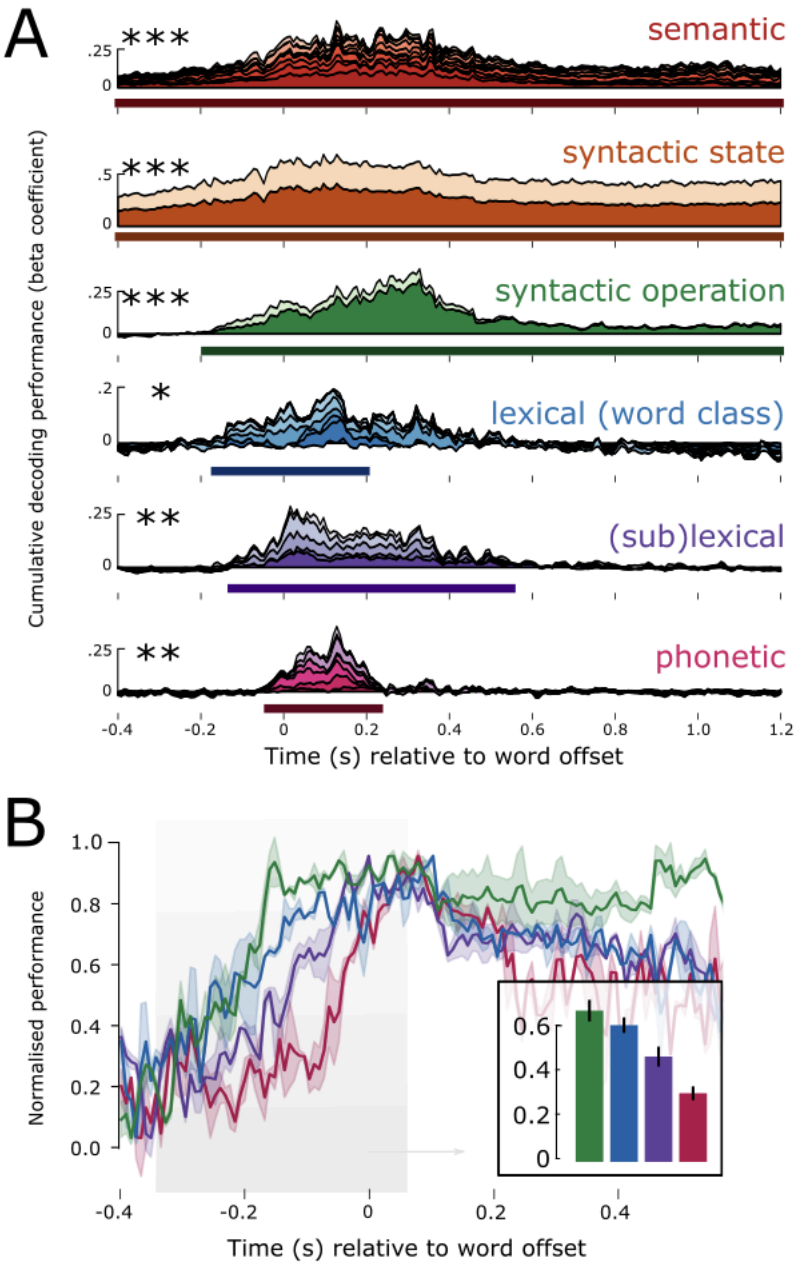

Figure 4: Decoding hierarchical features. A: Result of decoding each language level over time. The beta coefficients of each feature are stacked on top of each other, such that the top of the timecourse plot corresponds to the cumulative sum of all features in that linguistic level. The x-axis corresponds to time in seconds relative to word offset. The y-axis corresponds to the cumulative beta-coefficient across features. Solid line below the time-course represents the extent of the significant temporal cluster; asterisks represent its significance. B: Decoding performance zooming in for the lowest four feature families, and showing the standard error across subjects. Higher level features come online earlier than lower level ones. This is shown in the barplot, averaging performance before word offset shows a linear decrease in amplitude. * p < .05; ** p < .01; *** p < .001

First, we assessed the relationship between hierarchy and decoding onset time. For this, we normalized the decoding performance for each feature family, by dividing the group average by its maximum, for each family separately. We analyzed the rise-time before word offset (for the analysis on word onset see Supplementary Figures). As shown in Figure 4B, higher-level features were detectable earlier than lower-level features, resulting in a significant negative correlation between hierarchical level and the peak of the normalized performance (r = -0.82, p < .001).

Second, we tested the relationship between hierarchy and decoding duration. We found that higher-level features were decodable significantly longer than lower-level features, resulting in a significant positive correlation between level and duration (r = +0.75, p < .001). This effect was particularly striking for the syntactic and semantic features, which were decodable for over 1s after word offset, continuing well into the processing of the subsequent words (see Supplementary Figure 5 for the distribution of latencies of upcoming words).

Third, we tested the extent to which different features of the hierarchy are represented in parallel. We found evidence of a nested temporal structure, whereby the decodable window of a given level (L) was generally contained within the decodable window of the feature at L+1. For example, the start and end of significant phoneme decoding falls within the start and end of (sub)lexical decoding, and that in turn within the start and end of word class decoding, etc. A one-way F-test revealed that the entire hierarchy as defined by the 6 feature families was decodable in parallel from -40:230 ms (F-value in the cluster) = 4.1, p < .001) relative to word offset, i.e., throughout the duration of phonetic processing of the final speech sound of the word.

Together, these results confirm a key prediction of HPC: the dynamics of processing are increasingly sustained as the feature under consideration is high in the language hierarchy. We also observe that information at each level is encoded well into the processing of subsequent phonemes and words, leading to significant parallel processing, across and within levels of representation.

1.3. The neural bases of language exhibit hierarchical dynamics.

Overall, finding that each linguistic feature can be decoded – and is thus represented – for a longer time window than its actual duration in natural speech raises a critical question: how do successive representations avoid mutual interference?

Here, we test the Hierarchical Coding Hypothesis: a dynamic neural code allows successive phonemes, syllables and words to be maintained without interfering with one another. To test this, we implemented a temporal generalization analysis [4,20] (Figure 1). This method involves evaluating whether the topographic pattern learnt at time t generalizes to subsequent and preceding time-points (see Methods for details). If the representation is held within the same neural population over time, then the topographic pattern learnt at time t should generalize to time t+N, leading to a ‘square’ decoding matrix. By contrast, if the neural code evolves as a function of time, then the topographic pattern learnt at time t would not be the same at time t+N, even if the representation can also be decoded at t+N. In this scenario of a dynamic code, we thus expect to detect a ‘diagonal’ matrix. In sum, we can compare two durations: (1) the window during which the representation can be decoded and (2) the window during which decoders tend to generalize.

We applied this analysis to each of our language features, and then averaged the generalization matrices over the six levels of interest (Figure 5) to estimate the similarity of spatial evolution across the hierarchy.

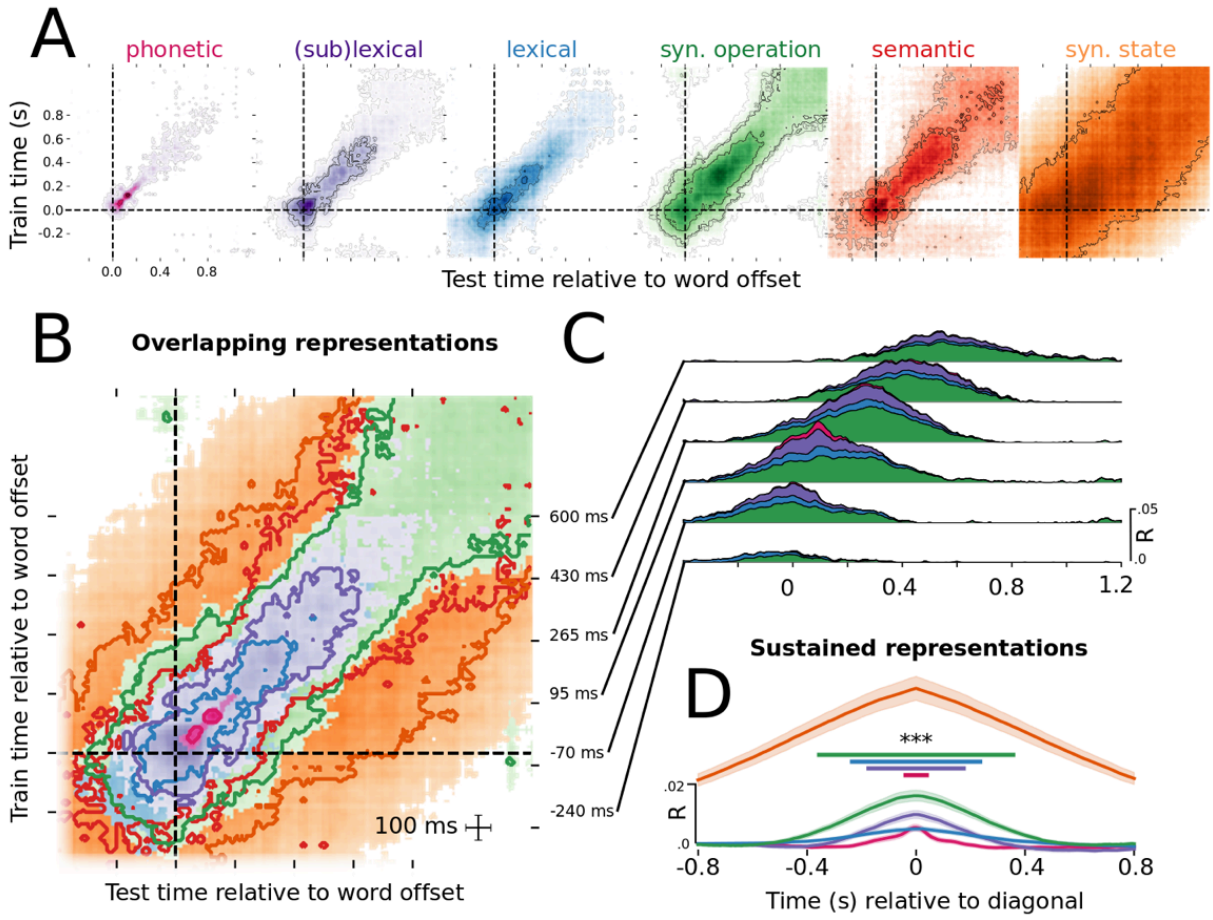

Figure 5: Evidence for Hierarchical Dynamic Coding. A: Temporal generalization analysis for each of the six linguistic levels of analysis. Each contour represents a significance threshold of p < .05, p < .01, p < .005, and p < .001. B: Same data as shown in A but for just the p < .01 threshold. C: Cumulative temporal generalization performance for the temporal decoders trained at different time-points relative to word offset. D: Data re-aligned relative to the diagonal of the temporal generalization matrix, showing the relationship between format maintenance and feature complexity. ***: *p* < .001.

We found that all six feature families are processed using a dynamic neural code. The neural activity patterns associated with each linguistic features are only stable for a relatively short time period: phonetic duration=184 ms; sustain=64 ms; sub-lexical duration=752 ms; sustain=384 ms; word class duration=536ms; sustain=224 ms; syntactic operation duration=1392ms; sustain=720 ms; syntactic state duration=1250 ms; sustain=1600 ms). This means that all levels of representation across the hierarchy are supported by neural populations that change over time.

Furthermore, the stability of a linguistic feature depends on its level in the language hierarchy: The lower-level phonetic features, which are defined over smaller linguistic units (phonemes), evolved significantly faster (average generalization time 64 ms), than lexical features (224 ms), and those, faster than syntactic features (730 ms). This led to a significant correlation between the location of the family in the hierarchy, and duration of information sustain (r = -0.89, p = .034) (Figure 5D). This finding suggests that while all levels of the hierarchy share a dynamic coding scheme, the speed with which information is routed to different neural populations scales with unit duration and abstraction.

Discussion

The present study shows that Hierarchical Dynamic Coding (HDC) provides a succinct account of how the brain continuously coordinates the hierarchy of representations underlying speech comprehension. Specifically, we find, with a comprehensive suite of 54 linguistic features decoded from MEG, that language representations are systematically embedded in a neural code, which changes at a speed commensurate with the corresponding linguistic level.

1. Content: Hierarchical representations

We decoded 54 speech features, organized within six levels of representation. These features are motivated by linguistic theory and are claimed to be relevant dimensions of language structure and content. However, their role in language processing remains contentious [1,24]. Our study thus clarifies the validity of these theories for understanding how the brain processes language. First, we find that MEG responses reliably track three highly-debated sublexical units: phonemes, syllables, and morphemes (Figure 3) [8,25]. This result strengthens recent fMRI findings [26], and suggests that these units may, in fact co-exist in the brain, and thus *all* serve as linguistic primitives. In other words, the brain may not have any quandary with generating multiple, partially redundant representations from the speech input. Rather, it can use these ubiquitous representations depending on the required task (e.g., speech segmentation; word-formation; grammatical analysis).

Second, our decoders reveal the timing of multiple syntactic features. Although syntax is often considered a landmark component of language [27–29], its neural underpinnings are notoriously challenging to study [10,12,13,30,31]. Here, we show that the brain represents grammatical properties both at the word level (e.g. part of speech) and at the tree-structure level (e.g. the complexity of the syntactic tree). Furthermore, these two sub- and supra-lexical levels appear to be linked by detectable syntactic operations, namely node opening and node closing. Although much in-depth work remains necessary, the possibility to track such symbolic operations pave the way to understand how such symbolic structures are built by neuronal populations.

2. Code: Hierarchical dynamic coding (HDC)

The existence of HDC is here characterized by three main findings. First, all levels of the language hierarchy are embedded in a continuously-evolving neural code. The “diagonal” temporal generalizations we observe (Figure 5) indicate that the very same linguistic feature can be decoded from distinct neural patterns across time. This means that each feature is represented by a sequence of neural populations rather than being maintained within a static activity pattern [20,21]. Importantly, this dynamic coding is present, not just for one level of representation as previously reported [4,23] but across all levels of the hierarchy. This result thus indicates a canonical processing strategy independent of feature abstraction and timescale. Overall, HDC effectively resolves the tension between the two constraints of speech comprehension: namely (1) maintaining individual linguistic representations, while (2) continuously updating them as new speech inputs are received.

Second, higher level features are maintained significantly longer in the neural signal, extending into the processing of multiple words in the future (Figure 4). This shows that each feature is maintained much longer than their corresponding presence in the speech input, and, consequently, successive phonemes, syllables, words, and syntactic states are simultaneously represented at each time point. This temporal overlap is much longer than previously appreciated [3,10].

Third, the speed of neural dynamics reduces with the level of language hierarchy: The neural code changes less rapidly for high level representations than for low level representations, as marked by different “diagonal widths” in the temporal generalization analysis (Figure 5). This finding echoes a large body of literature on the oscillatory bases of language processing, whereby higher-level representations (e.g. words) are linked to slower oscillations (e.g. delta) than lower-level representations (e.g. syllables = theta) [32–34]. Anatomically, these results also fit with the fact that higher-level areas (e.g. fronto-parietal

cortices) integrate speech representations over longer time periods than lower-level areas (e.g. Heschl gyrus) [35–39]. Our results complement these studies, by showing that this hierarchy reflects the necessity to both maintain successive features and avoid their representational interference at each level of the hierarchy.

3. Bridging the explanatory gap between AI and the neuroscience of language

Our approach differs from alternative methods based on the language features automatically generated by large language models [37,40–43]. While powerful in accounting for brain responses, the resulting features are notoriously difficult to interpret. One interesting point of potential convergence with large language models, however, is their use of the 'transformer' architecture, which implements "positional embedding" [44,45]: i.e., a continuously evolving projection which helps the model to temporally locate each word in a sentence. Such position embeddings apply a geometrically-predictable transformation to the representations of successive tokens. To some extent, we observe a similar process, whereby each hierarchical representation is projected by a dynamical embedding, which preserves both the order and the fidelity of the incoming information. Beyond comparing the representations of deep nets and those of the brain, it will be important, in the future, to evaluate whether their coding schemes may indeed converge.

4. Evidence of a reverse hierarchy

An important contribution of our work is in linking different levels of the language hierarchy to distinct processing dynamics. One striking observation is the inverse relationship between feature complexity and onset of significant feature decoding relative to word offset. Higher-level features (syntactic properties) come online ~500 ms before low-level features (phonetic and sub-lexical properties) (Figure 3). This result is in line with the 'reverse hierarchy' theory as pioneered in vision [46–48], which stipulates that high-level representations (e.g., object identity: it is a cat; it is a house) are immediately activated based on prediction from context, and thus offer immediate access to global information that forms the *gist* of the visual scene. Only later are local details of the input filled in (e.g., orientation, color). A similar process could occur in speech processing: Higher order structures such as syntactic frames and semantic contents may be immediately available based on the context, and thus serve to guide the interpretation of lower levels, such as lexical identity and phonological form. In this view, this reverse hierarchy would only be observable when the language input has a context which provides an overarching semantic topic and syntactic frame. Without this high-level information, processing necessarily proceeds in a compositional, bottom-up fashion, because there is no context upon which to bootstrap subsequent processes. Because previous studies often present subjects with de-contextualized language input (i.e. isolated words or phrases), thus withholding the context to construct higher-order information, prior work may have been unable to observe the phenomenon we describe here.

A reverse hierarchy architecture for speech processing presents three computational advantages. First, because higher-level language features are abstracted away from the sensory signal, comprehension would be more robust to auditory noise and ambiguity [49,50]. Second, restricting the search space of possible words from the highest levels means that predictions are formed based on the syntactic and semantic content of the message, which are more robust against acoustic noise, and uninformative phonetic variation [51]. Finally, it potentially speeds up processing by initiating high-level computations early during comprehension, rather than waiting for them to be formed compositionally in purely bottom-up fashion [52–54].

Overall, this suggests that in continuous speech, processing does not unfold from the bottom-up. Rather, higher order semantic and syntactic structures are available to the processing system earlier than the sensory signal of the speech sounds being processed.

5. Conclusion

Overall, our results show that the brain generates a sophisticated – yet interpretable – set of linguistic features to comprehend speech. These features, ranging from phonemes to syntactic trees, follow

canonical coding dynamics, which adapts its processing speed as a function of level in the language hierarchy. The resulting Hierarchical Dynamic Coding (HDC) elegantly balances the preservation of information over time with minimizing overlap between consecutive language elements. This system provides a clear view of how the brain may organize and interpret speech in real time, linking linguistic theories with their neurological foundations.

Acknowledgements

We thank Graham Flick for help with data collection. A big thanks to Florencia Assaneo, Joan Opella, Arianna Zuanazzi, Suzanne Dikker and Jill Kries for feedback on a previous version of the manuscript. **Funding**: This project received funding from the Abu Dhabi Institute G1001 (AM); NIH R01DC05660 (DP), European Union's Horizon 2020 research and innovation program under grant agreement No 660086, the Bettencourt-Schueller Foundation, the Fondation Roger de Spoelberch, the Philippe Foundation, the FrontCog grant ANR-17-EURE-0017 to JRK for his work at NYU and PSL and The William Orr Dingwall Dissertation Fellowship (LG). **Author contributions**: LG: conceptualisation; methodology; software; validation; formal analysis; investigation; data curation; writing - original draft preparation and review and editing; visualization. JRK: conceptualisation; methodology; software; supervision. AM: conceptualisation; writing - review and editing; supervision; funding acquisition. DP: conceptualisation; writing - review and editing; supervision; funding acquisition. **Competing interests**: The authors declare no competing interests. **Data and materials availability**: Preprocessed data have been publicly released on the Open Science Framework (https://osf.io/ag3kj/, [55]).

Methods

3.0. Definition of terms:

- **Feature**: a property of language: e.g., “fricative” phonetic feature; word frequency, etc.
- **Level**: a group of features at the same degree of hierarchical position: e.g. acoustic (pre symbolic), sub-lexical, lexical, supra lexical
- **Representation**: neural encoding scheme of a feature
- **Dynamic coding**: when neural representations are instantiated by distinct neural populations over time.
- **Hierarchical Dynamic Coding**: when different levels follow a dynamic coding scheme

3.1. Participants

Twenty-one native English participants were recruited from the NYU Abu Dhabi community (13 female; age: M=24.8, SD=6.4). All provided their informed consent and were compensated for their time. Participants reported having normal hearing and no history of neurological disorders. Each subject participated in the experiment twice. Time between sessions ranged from 1 day to 2 months. All participants gave their informed consent, and the experiment was approved by the local IRB committee of NYU Abu Dhabi.

3.2. Stimulus development

Four fictional stories were selected from the Open American National Corpus [56]: Cable spool boy (about two bothers playing in the woods); LW1 (sci-fi story about an alien spaceship trying to find home); Black willow (about an author struggling with writer’s block); Easy money (about two old friends using magic to make money).

Stimuli were annotated for phoneme boundaries and labels using the ‘gentle aligner’ from the Python module *lowerquality*. Some prior testing provided better results than the Penn Forced Aligner [57].

Each of the stories were synthesised using the Mac OSX text-to-speech application. Three synthetic voices were used (Ava, Samantha, Allison). Voices changed every 5-20 sentences. The speech rate of the voices ranged from 145-205 words per minute, which also changed every 5-20 sentences. The silence between sentences randomly varied between 0-1000 ms.

3.3. Procedure

Before the experiment proper, the participant was exposed to 20 seconds of each speaker explaining the structure of the experiment. This was designed to help the participants attune to the synthetic voices. The order of stories was fully crossed using a Latin-square design. Participants heard the stories in the same order during both the first and second sessions. This was in order to make direct comparisons between the first and second sessions.

Participants answered a two-choice question on the story content every ~3 minutes. For example, one of the questions was “what was the location of the bank that they robbed”? The purpose of the questions was to keep participants attentive and to have a formal measure of engagement. All participants performed this task at ceiling, with an accuracy of 98%. Participants responded with a button press. Stimuli were presented binaurally to participants though tube earphones (Aero Technologies), at a mean level of 70 dB SPL. The stories ranged from 8-25 minutes, with a total running time of ~1 hour.

3.4. MEG acquisition

Marker coils were placed at the same five positions to localise each participant's skull relative to the sensors. These marker measurements were recorded just before and after the experiment in order to track the degree of movement during the recording.

MEG data were recorded continuously using a 208 channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan), with a sampling rate of 1000 Hz and applying an on- line low-pass filter of 200 Hz.

3.5. Preprocessing MEG

The raw MEG data were noise reduced using the Continuously Adjusted Least Squares Method (CALM: [58], with MEG160 software (Yokohawa Electric Corporation and Eagle Technology Corporation, Tokyo, Japan).

The data were bandpass-filtered between 0.1 and 50 Hz using MNE-Python's default parameters with firwin design [50] and downsampled to 250 Hz. Epochs were segmented from 200 ms pre-phoneme onset to 600 ms post-phoneme onset. No baseline correction was applied.

3.6. Effects of acoustic features

We used a temporal receptive field (TRF) model to regress from the raw MEG data responses that were sensitive to fluctuations in the pitch and envelope of the acoustic speech signal. We used the *ReceptiveField* function from MNE-Python [59], using ridge regression as the estimator and laplacian regularization. We tested ten lambda regularization parameters, log-spaced between 1^{-6} and 1^{+6}, and picked the model with the highest predictive performance averaged across sensors. MEG sensor activity at each ms were modeled using the preceding 200 ms of envelope and pitch estimates. Both the acoustic and MEG signals were demeaned and scaled to have unit variance before fitting the model. MEG acoustic-based predictions were then transformed back into original MEG units before regressing out of the true MEG signals. This process, including fitting hyper-parameters, was applied for each story recording and for each subject separately, across 3 folds. This yields a de-confounded MEG dataset on which to continue our analysis.

3.7. Modeled features

We investigated whether single-trial sensor responses varied as a function of 54 features. Features spanned different levels of the linguistic hierarchy and included both binary and continuous variables.

3.7.1. Phoneme-level features

Phonetic features were derived from the multi-value feature system reported in [60]. Note that this feature system is sparse relative to the full set of distinctive features that can be identified in English; however, it serves as a reasonable approximation of the phonemic inventory for our purposes.

Voicing. This refers to whether the vocal chords vibrate during production. For example, this is the difference between *b* versus *p* and *z* versus *s*.

Manner of articulation. Manner refers to the way by which air is allowed to pass through the articulators during production. Here we tested five manner features: fricative, nasal, plosive, approximant, and vowel.

Place of articulation. Place refers to where the articulators (teeth, tongue, lips) are positioned during production. For vowels, this consists of: central vowel, low vowel, mid vowel, high vowel. For consonants, this consists of: coronal, glottal, labial and velar.

3.7.2. Sub-lexical features

Number of phonemes, syllables and morphemes in the word. These features were obtained from the English Lexicon Project (ELP) [61], which included a corpus of spoken English. They constitute a count of the total number of units within the word.

Number of phonemes in the syllable. Also derived from the ELP. Constitutes a count of the total number of phonemes within the current syllable.

Word frequency. Also derived from the ELP. Corresponds to the log frequency of the word from the subtitles corpus of American English.

Phonological neighbourhood density. Also derived from the ELP. Corresponds to how many other words can be formed by substituting a single phoneme in the word to create another word. This has been reported as a metric of lexical competition.

3.7.3. Word class

All word class labels were derived from the syntactic parse of the stories. They are dummy coded relative to 11 word-class categories: Adjective (e.g. blue, tall); Coordinating conjunction (e.g. and, or); Determiner (e.g. the, a); Noun (e.g. house; girl); Pronoun (e.g. she, they); Preposition (e.g. under, on); Adverb (e.g. slowly, fast); Verbal preposition (e.g. to); Verb (e.g. run, jump); WH-Word (e.g. where, who); Existential there (e.g. there).

3.7.4. Syntactic operation

These features are derived from the penn-treebank syntactic parse of the stories. Number of closing nodes corresponds to how many brackets of the tree are being closed at a particular word; opening nodes corresponds to how many brackets are opened at a particular word. Sentence end corresponds to a binary coding of whether the word is in sentence-final position.

3.7.5. Syntactic state

These features are derived from the penn-treebank syntactic parse of the stories. Number of open nodes corresponds to how many brackets remain open at a particular word. Tree depth corresponds to the number of levels down a tree a particular word is situated. Tree depth +/- 1 corresponds to the depth at the immediately preceding and subsequent word. These regressors are included to account for auto-correlation in the stimulus features. Linear order corresponds to the location of the word in the sentence; +/- corresponds to the same at the previous and next word.

3.7.6. Semantic vector

We obtained the 50-dimensional word embedding GloVe vectors [61,62] for each of our words and applied a principal component analysis, yielding ten principal components. These components are organised relative to the amount of variance in the original GloVe embeddings they correspond to.

3.8. Back-to-back regression decoding

We fit a back-to-back regression algorithm [63], which allows us to decode multiple features from the MEG data while also controlling for their co-variation. The resulting model coefficients represent how robustly a linguistic feature is encoded in neural responses, above and beyond the variance accounted for by the other linguistic features.

For the neural decoding, the input features were the magnitude of activity at each of the 208 MEG sensors. This approach allows us to decode from multiple, potentially overlapping, neural representations, without relying on gross modulations in activation strength [63,64].

Because some of the features in our analysis are correlated with one another, we need to jointly evaluate the accuracy of each decoding model relative to its performance in predicting all modeled features, not just the target feature of interest. This is because, if fitting each feature independently, we will not be able to dissociate the decoding of feature *f* from the decoding of the correlated feature *f*̂. The necessity to use decoding over encoding models here, though (which, do not suffer so harshly from the problem of co-variance in the stimulus space) is one of signal to noise: we expect any signal related to linguistic processes to be contained in low-amplitude responses that are distributed over multiple sensors. Our chances of uncovering reliable responses to these features is boosted by using multivariate models.
To overcome the issue of covariance, but still to capitalize on the advantages of decoding approaches, we implement a back-to-back ridge regression model [30]. This involves a two stage process. First, a ridge regression model was fit on a random half of the data, at a single time- point. The mapping was learnt between the multivariate input (activity across sensors) and the univariate stimulus feature (one of the 54 features described above). All decoders were provided with data normalized by the mean and standard deviation in the training set:

$$arg\,min\beta \sum_i (y_i \beta^T X_i)^2 + \alpha \|\beta\|^2$$

where $y_i \in \{\pm 1\}$ is the feature to be decoded at trial *i* and X_i is the multivariate neural measure.
The l2 regularization parameter α was also fit, testing 20 log-spaced values from 1^{-5} to 1^5. This was implemented using the *RidgeCV* function in *scikit-learn* [65].

Then, we use the other half of the acoustic or neural responses to generate a prediction for each of the 31 features corresponding to the test set. However, because the predictions are correlated, we need to jointly-evaluate the accuracy of decoding each feature, to take into account the variance explained by correlated non-target features. To do this, we fit another ridge regression model, this time learning the beta coefficients that map the matrix of *true* feature values to *predicted* feature values:

$$arg\,min\beta \sum_i (y_i \beta^T \hat{Y}_i)^2 + \alpha \|\beta\|^2$$

where $y_i \in \{\pm 1\}$ is the ground truth of a particular stimulus feature at trial *i* and $\hat{Y}_i$ is the prediction for all stimulus features. A new regularization parameter α was learnt for this stage. By including all stimulus features in the model, this accounts for the correlation between the feature of interest and the other features. From this, we use the beta-coefficients that maps the true stimulus feature to the predicted stimulus feature. Beta coefficients serve as our metric of decoding performance: if a stimulus features is not encoded in neural responses (the null hypothesis) then there will be no meaningful mapping between the true feature *y* and the model prediction *y*̂. Thus, the beta coefficient will be zero – equivalent to chance performance. If, however, a feature *is* encoded in neural activity (the alternative hypothesis), we should uncover a significant relationship between *y* and *y*̂, thus yielding an above-zero beta coefficient.

The train/test split was performed over 100 folds, and the beta-coefficients were averaged across folds. This circumvents the issue of unstable coefficients when modeling correlated variables. These steps were applied to each subject independently.

3.9. Temporal generalization decoding

Temporal generalization (TG) consists of testing whether a temporal decoder fit on a training set at time *t* can decode a testing set at time *t'* [20]. This means that rather than evaluating decoding accuracy just at the time sample that the model was trained on, we evaluate its accuracy across all possible train/testing time combinations.

TG can be summarized with a square training time × testing time decoding matrix. To quantify the stability of neural representations, we measured the duration of above-chance generalization of each temporal

decoder. To quantify the dynamics of neural representations, we compared the mean duration of above-chance generalization across temporal decoders to the duration of above-chance temporal decoding (i.e. the diagonal of the matrix versus its rows). These two metrics were assessed within each subject and tested with second-level statistics across subjects.

3.10. Comparing decoding performance between trial subsets

To evaluate whether the processing of syntax built over time, we subset our analysis by evaluating decoding performance at different word positions in the sentence. We add a modification to our train/test cross-validation loop. The data are trained on the entire training set (i.e. the same number of trials as the 'typical analysis'), and the test set is grouped into the different levels of interest. We evaluate model performance separately on each split of the test data, which yields a time-course or generalization matrix for each group of trials that we evaluate on: in this case, each word position in the sentence.

3.11. Group statistics

To evaluate whether decoding performance is better than chance, we perform second- order statistics. This involves testing whether the distribution of beta coefficients across subjects significantly differs from chance (zero) across time using a one-sample permutation cluster test with default parameters specified in the MNE-Python package [59].

References

1. Halle, M. & Stevens, K. Speech recognition: A model and a program for research. *IRE Transactions on Information Theory* **8**, 155–159 (1962).
2. Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. Perception of the speech code. *Psychol. Rev.* **74**, 431–461 (1967).
3. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
4. Gwilliams, L., King, J.-R., Marantz, A. & Poeppel, D. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nat. Commun.* **13**, 6606 (2022).
5. Oganian, Y. & Chang, E. F. A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci Adv* **5**, eaay6279 (2019).
6. Poeppel, D. & Assaneo, M. F. Speech rhythms and their neural foundations. *Nat. Rev. Neurosci.* **21**, 322–334 (2020).
7. Gwilliams, L., Linzen, T., Poeppel, D. & Marantz, A. In Spoken Word Recognition, the Future Predicts the Past. *J. Neurosci.* **38**, 7585–7599 (2018).
8. Gwilliams, L. How the brain composes morphemes into meaning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190311 (2020).
9. Keshishian, M. *et al.* Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nat Hum Behav* **7**, 740–753 (2023).
10. Bemis, D. K. & Pylkkänen, L. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J. Neurosci.* **31**, 2801–2814 (2011).
11. Pallier, C., Devauchelle, A.-D. & Dehaene, S. Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2522–2527 (2011).
12. Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M. & Hale, J. T. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain Lang.* **157-158**, 81–94 (2016).
13. Nelson, M. J. *et al.* Neurophysiological dynamics of phrase-structure building during sentence

processing. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3669–E3678 (2017).

14. Hickok, G. & Poeppel, D. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* **92**, 67–99 (2004).
15. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
16. Leonard, M. K. & Chang, E. F. Dynamic speech representations in the human temporal lobe. *Trends Cogn. Sci.* **18**, 472–479 (2014).
17. Honey, C. J. *et al.* Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).
18. Fedorenko, E. & Thompson-Schill, S. L. Reworking the language network. *Trends Cogn. Sci.* **18**, 120–126 (2014).
19. Desbordes, T. *et al.* Dimensionality and Ramping: Signatures of Sentence Integration in the Dynamics of Brains and Deep Language Models. *J. Neurosci.* **43**, 5350–5364 (2023).
20. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).
21. Stokes, M. G., Buschman, T. J. & Miller, E. K. Dynamic coding for flexible cognitive control. in *The Wiley Handbook of Cognitive Control* 221–241 (John Wiley & Sons, Ltd, 2017). doi:10.1002/9781118920497.ch13.
22. Stroud, J. P., Watanabe, K., Suzuki, T., Stokes, M. G. & Lengyel, M. Optimal information loading into working memory explains dynamic coding in the prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2307991120 (2023).
23. King, J.-R. & Wyart, V. The Human Brain Encodes a Chronicle of Visual Events at Each Instant of Time Through the Multiplexing of Traveling Waves. *J. Neurosci.* **41**, 7224–7233 (2021).
24. Halle, M. & Marantz, A. Distributed morphology and the pieces of inflection. *1993* 111–176 (1993).
25. Poeppel, D. The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. *Speech Commun.* **41**, 245–255 (2003).
26. Gong, X. L. *et al.* Phonemic segmentation of narrative speech in human cerebral cortex. *Nat. Commun.* **14**, 4309 (2023).

27. Van Valin, R. D. *An Introduction to Syntax*. (Cambridge University Press, 2001).
28. Chomsky, N. *Topics in the Theory of Generative Grammar*. (Walter de Gruyter, 2013).
29. Chomsky, N. *Aspects of the Theory of Syntax, 50th Anniversary Edition*. (MIT Press, 2014).
30. Zioga, I., Weissbart, H., Lewis, A. G., Haegens, S. & Martin, A. E. Naturalistic Spoken Language Comprehension Is Supported by Alpha and Beta Oscillations. *J. Neurosci.* **43**, 3718–3732 (2023).
31. Hale, J. T. *et al.* Neurocomputational Models of Language Processing. *Annu. Rev. Linguist.* **8**, 427–446 (2022).
32. Giraud, A.-L. & Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
33. Doelling, K. B., Arnal, L. H., Ghitza, O. & Poeppel, D. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* **85**, 761–768 (2014).
34. Morillon, B. & Schroeder, C. E. Neuronal oscillations as a mechanistic substrate of auditory temporal prediction. *Ann. N. Y. Acad. Sci.* **1337**, 26–31 (2015).
35. 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).
36. Caucheteux, C., Gramfort, A. & King, J.-R. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. *arXiv [q-bio.NC]* (2021).
37. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat Hum Behav* **7**, 430–441 (2023).
38. Jain, S., Vo, V. A., Wehbe, L. & Huth, A. G. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language* 1–65 (2023).
39. Chang, C. H. C., Nastase, S. A. & Hasson, U. Information flow across the cortical timescale hierarchy during narrative construction. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2209307119 (2022).
40. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
41. Caucheteux, C., Gramfort, A. & King, J.-R. Disentangling syntax and semantics in the brain with deep networks. in *Proceedings of the 38th International Conference on Machine Learning* (eds.

Meila, M. & Zhang, T.) vol. 139 1336–1348 (PMLR, 18--24 Jul 2021).

42. D'efossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.* **5**, 1097–1107 (2022).
43. Vaidya, A. R., Jain, S. & Huth, A. G. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv [cs.CL]* (2022).
44. Su, J. *et al.* RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing* **568**, 127063 (2024).
45. Peng, B., Quesnelle, J., Fan, H. & Shippole, E. YaRN: Efficient Context Window Extension of Large Language Models. *arXiv [cs.CL]* (2023).
46. Hochstein, S. & Ahissar, M. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36**, 791–804 (2002).
47. Ahissar, M. & Hochstein, S. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* **8**, 457–464 (2004).
48. Ahissar, M., Nahum, M., Nelken, I. & Hochstein, S. Reverse hierarchies and sensory learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 285–299 (2009).
49. Warren, R. M. Perceptual restoration of missing speech sounds. *Science* **167**, 392–393 (1970).
50. Leonard, M. K., Baud, M. O., Sjerps, M. J. & Chang, E. F. Perceptual restoration of masked speech in human cortex. *Nat. Commun.* **7**, 13619 (2016).
51. Nahum, M., Nelken, I. & Ahissar, M. Low-level information and high-level perception: the case of speech in noise. *PLoS Biol.* **6**, e126 (2008).
52. Friederici, A. D. Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* **6**, 78–84 (2002).
53. Ferreira, F., Bailey, K. G. D. & Ferraro, V. Good-Enough Representations in Language Comprehension. *Curr. Dir. Psychol. Sci.* **11**, 11–15 (2002).
54. Frances, C. Good enough processing: what have we learned in the 20 years since Ferreira et al. (2002)? *Front. Psychol.* **15**, (2024).
55. Gwilliams, L. *et al.* Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Sci Data* **10**, 862 (2023).

56. Ide, N. & Macleod, C. The american national corpus: A standardized resource of american english. in *Proceedings of corpus linguistics* vol. 3 1–7 (Lancaster University Centre for Computer Corpus Research on Language …, 2001).

57. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123**, 3878–3878 (2008).

58. Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y. & Ochiai, M. Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. *IEEE Trans. Appl. Supercond.* **11**, 669–672 (2001).

59. Gramfort, A. *et al.* MNE software for processing MEG and EEG data. *Neuroimage* **86**, 446–460 (2014).

60. King, S. & Taylor, P. Detection of phonological features in continuous speech using neural networks. *Comput. Speech Lang.* **14**, 333–353 (2000).

61. Balota, D. A. *et al.* The English Lexicon Project. *Behav. Res. Methods* **39**, 445–459 (2007).

62. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2014). doi:10.3115/v1/d14-1162.

63. King, J.-R., Charton, F., Lopez-Paz, D. & Oquab, M. Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *Neuroimage* **220**, 117028 (2020).

64. King, J.-R., Gramfort, A. & Others. Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition. (2018).

65. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* 2825–2830 (2012).