Title: Predicting speech from a cortical hierarchy of event-based timescales

Short title: Surprisal hierarchy in speech prediction

Authors: Lea-Maria Schmitt[1,2*], Julia Erb[1,2], Sarah Tune[1,2], Anna Rysop[3], Gesa Hartwigsen[3], Jonas Obleser[1,2*]

Affiliations:

[1] Department of Psychology, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany.

[2] Center of Brain, Behavior and Metabolism, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany.

[3] Lise Meitner Research Group Cognition and Plasticity, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1 A, 04103 Leipzig, Germany.

* Corresponding authors. Email: *l.schmitt@uni-luebeck.de*; *jonas.obleser@uni-luebeck.de*.

Abstract: How is the temporal unfolding of context in our natural environment processed in the brain to inform predictions? We here provide evidence for a neural coding scheme that sparsely updates contextual representations at the boundary of events and gives rise to a hierarchical, multi-layered organization of predictive language comprehension. Training artificial neural networks to predict the next word in a story at five stacked timescales and then using model-based functional MRI, we observe a sparse, event-based "surprisal hierarchy". This hierarchy evolved along a temporo-parietal pathway, with surprisal at longest timescales represented in inferior parietal regions. Along this hierarchy, surprisal at any given timescale gated bottom-up and top-down connectivity to neighbouring timescales. In contrast, surprisal derived from continuously updated context influenced temporo-parietal activity only at short timescales. Representing context in the form of increasingly coarse events constitutes a network architecture for making predictions that is both computationally efficient and contextually diverse.

Teaser: A temporo-parietal pathway structures human speech into hierarchically organized events to make predictions on upcoming words

Introduction

While the past predicts the future, not all context the past provides is equally informative: it might be outdated, contradictory, or even irrelevant. Nevertheless, the brain as a "prediction machine" (*1*) is seemingly equipped with a versatile repertoire of computations to overcome these contextual ambiguities. A prominent example is speech, where a slip of the tongue may render the most recent context uninformative, but we can still predict the next word from its remaining context. At much longer time scales, we can re-use context that suddenly proves informative, as a speaker returns to a topic discussed earlier.

Using natural language comprehension as a working model, we here ask: How does the brain dynamically organize, evaluate, and update these complex contextual dependencies over time to make accurate predictions?

A robust principle in cerebral cortex is the decomposition of temporal context into its constituent timescales along a hierarchy from lower to higher-order areas, which is evident across species (*2*, *3*), recording modalities (*4*, *5*), sensory modalities (*6*, *7*), and cognitive functions (*8*, *9*). For instance, sensory cortices closely track rapid fluctuations of stimulus features and operate on short timescales (e.g., Ref. (*10*)). By contrast, association cortices integrate stimuli over an extended period and operate on longer timescales (e.g., Ref. (*11*)). Indeed, hierarchical specialization has been shown to emerge from structural and functional large-scale connectivity across cortex (*12*, *13*), with connectivity between neighbouring areas allowing for efficient mapping between timescales (*14*, *15*).

Conceptually, such hierarchies of "temporal receptive windows" are often subsumed under the framework of predictive coding (*16*): A nested set of timescale-specific generative models informs predictions on upcoming sensory input and is updated based on the actual input (*17*). In particular, context is thought to shape the prediction of incoming stimuli via feedback connections. These connections would link each timescale to its immediate shorter timescale, while the prediction error is propagated forward through the hierarchy (*18*). In line with such accounts of hierarchical predictive coding, feedforward and feedback connections in cortex (*19*, *20*) have been shown to carry prediction errors and predictions (*21*, *22*), respectively.

However, studies on the neural underpinnings of predictive coding have primarily used artificial stimuli of short temporal context (but see Ref. (*23*)) and employed local versus global violations of expectations, effectively manifesting a two-level cortical hierarchy (but see Ref. (*24*)). We thus lack understanding whether the hierarchical organization of prediction processes extends to *natural* environments unfolding their temporal dependencies over a *multitude* of interrelated timescales.

With respect to functional organization in human cortex, temporo-parietal areas are sensitive to a rich set of hierarchies and timescales in speech (*25*, *26*). Most relevant to the present work, semantic context in a spoken story has been shown to map onto a gradient extending from early auditory

cortex representative of words up to intraparietal sulcus representative of paragraphs (*27*) or the broader storyline in a narrative (*28*). This timescale-specific representation of context is reminiscent of the multi-layered generative models proposed to underlie predictive coding (*29*, *30*). Compatible with this notion, previous studies on speech comprehension found evidence for neural coding of prediction errors at the level of syllables (*31*), words (*32*), or discourse (*33*).

Yet the interactions between multiple representational timescales of speech in predicting upcoming words remain unclear. Here, we ask whether the temporo-parietal processing hierarchy enabling natural speech comprehension is also implicated in evaluating the predictiveness of timescale-specific semantic context and integrating informative context into predictions. We recorded blood-oxygen-level-dependent (BOLD) responses while participants listened to a narrated story, which provides rich semantic context and captures the full dynamic range of speech predictability. We set out to identify prediction error signalling, which can be formulated as surprisal (*34*). Psycholinguistic surprisal theory states that surprisal is proportional to the cognitive effort required to update a probabilistic prediction when encountering a new word (*35*, *36*), which modulates BOLD activity (*37*) and reading time (*38*).

Here, we focus on two potential computational architectures that could underlie a cortical hierarchy of predictive language comprehension. One attractively simple candidate architecture constructs context representations at all timescales with each upcoming word by replacing outdated context with context just recently encountered. An important implication of such a *continuously updating processing hierarchy* is that context representations at all timescales are tuned to current processing demands when predicting an upcoming word. In a recent study, Chien and Honey (*39*) showed that a computational model with continuous updates best explained how neural responses to a story rapidly aligned across participants in areas with shorter, but only later in areas with longer receptive windows.

A competing candidate architecture is effectively based on the event structure of context, with changes in situational features indicating the end of an event. For example, it is known that scenes in a movie are encoded as event-specific neural responses (*40*) and that more parietal receptive windows represent increasingly coarse events in movies (*41*). The segmentation of context into hierarchical events is implemented via an additional boundary detector that initiates the recombination of an event with preceding events at higher processing stages at an event boundary only. This sparsely updating processing hierarchy calls for fewer updates to context representations, which have been shown to speed up processing (*42*) and allow to draw on semantically more diverse context representations when making predictions. We here hypothesize that such a network architecture is a more appropriate model for prediction processes in the brain.

In the present study, we followed the rationale that neural computations can be inferred by comparing the fit of neural data to outputs from artificial neural networks with different

architectures (*43*, *44*). We derived context-specific surprisal associated with each word in the story from single layers of long short-term memory-based language models with either a continuous (LSTM (*45*)) or a sparse updating rule (hierarchical multiscale LSTM or HM-LSTM (*42*)). These language models make probabilistic predictions on upcoming words by exploiting representations of context at different timescales with the aim to minimize surprisal and the capability to learn from experience, thereby sharing key features with language prediction in humans (*30*). Further, LSTM models have been shown to effectively explain neural (*46*) and behavioural responses to speech (*47*).

Our results speak to the event-based organization of semantic context as a valid model of predictive processing in the brain. We show that a "surprisal hierarchy" of increasingly coarse event timescales evolves along the temporo-parietal pathway, with stronger connectivity to neighbouring timescales in states of higher word surprisal. Surprisal derived from continuously updated context had a (non-hierarchical) effect on temporo-parietal activity only at short timescales. Together, these results suggest that representing context in the form of hierarchical events constitutes a network architecture that is computationally efficient while providing a diverse range of context for making predictions.

114 # Results

115 Thirty-four participants listened to a one-hour narrated story while their hemodynamic brain
116 responses were recorded using functional magnetic resonance imaging (fMRI). To emulate a
117 challenging listening scenario, the story was presented against a competing stream of resynthesized
118 natural sounds (for an analysis focussing on cortical representations of acoustics see Ref. (*48*)).

119 The surprisal associated with each word in the story was modelled at multiple timescales of
120 semantic context by two artificial neural networks, an LSTM model with a continuous updating rule
121 and a HM-LSTM model with a sparse updating rule (Figure 1A).

122 First, we encoded surprisal at multiple timescales into univariate neural responses and fit a
123 gradient to temporo-parietal peak locations of timescales. Encoding models were estimated
124 separately per each language model, using ridge regression with four-fold cross-validation. Second,
125 we investigated how surprisal gates the information flow between brain regions sensitive to
126 different timescales.

127 **Two competing language models of hierarchical speech prediction**

128 We trained two artificial neural networks on more than 130 million words of running text to predict
129 an upcoming word by its preceding semantic context. More specifically, language models consisted
130 of LSTM cells (*45*), which incorporate context that might become relevant at *some* time (*cell state*) or
131 that is relevant already to the prediction of the *next* word (*hidden state*). By stacking five LSTM layers,
132 our models operated on different timescales of context, with higher layers coding for long-term
133 dependencies between words.

134 In the continuously updating (or "vanilla") LSTM, recurrent memory states are updated at each
135 layer with every new bottom-up word input (Figure 1B). A second model, the sparsely updating HM-
136 LSTM, employs a revised updating rule where information from a lower layer is only fed forward at
137 the end of an event (Figure 1C). This allows for less frequent updates between layers and stronger
138 separation between contextual information represented at different layers.

139 **Model-derived metrics of predictiveness at multiple timescales**

140 For each word in the entire presented story (> 9,000 words), we determined its predictability given
141 the semantic context of the preceding 500 words. Hidden states were combined across layers and
142 mapped to an output module, which denotes the probability of occurring next for every word in a
143 large vocabulary of candidate words. The word with the highest probability was selected as the
144 *predicted* next word. Overall, the LSTM (proportion correct across words: 0.13) and the HM-LSTM
145 (0.12; Figure S1) were on par in accurately predicting the next word in the story.

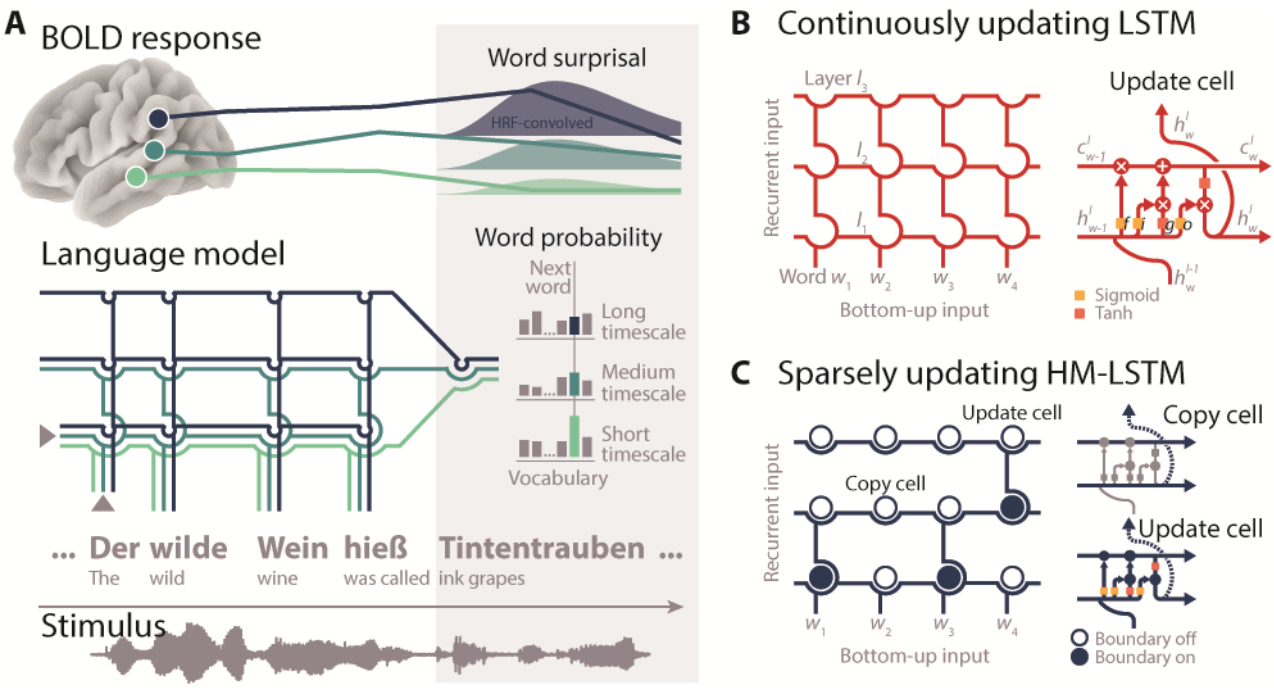

Figure 1. Modelling neural speech prediction with artificial neural networks. (A) Bottom: Participants listened to a story (grey waveform) during fMRI. Middle: Based on preceding semantic context ("The wild wine was called"), a language model assigned a probability of being the next word to each word in a large vocabulary (grey bars). The probability of the actual next word ("ink grapes"; coloured bars) was read out from each layer of the model separately, with higher layers (darker blue colours) accumulating information across longer semantic timescales. Top: Word probabilities were transformed to surprisal, convolved with the hemodynamic response function (bell shapes) and mapped onto temporo-parietal BOLD time series (coloured lines). **(B)** Two language models were trained. With each new word-level input, the "continuously updating" long short-term memory (LSTM (*45*)) combines "old" recurrent long-term (c^l_{w-1}) and short-term memory states (h^l_{w-1}) with "new" bottom-up linguistic input (h^{l-1}_w) at each layer *l*. This allowed information to continuously flow to higher layers with each incoming word. *f*: forget gate, *i*: input gate, *g*: candidate state, *o*: output gate. **(C)** The "sparsely updating" hierarchical multiscale LSTM (HM-LSTM (*42*)) was designed to learn the hierarchical structure of text. An upper layer keeps its representation of context unchanged (copy mechanism) until a boundary indicates the end of an event on the lower layer and information is passed to the upper layer (update mechanism). Networks were unrolled over the sequence of words for illustration only.

To derive the predictability of words based on layer-specific context (or, for our purpose, timescale), we allowed information to freely flow through pre-trained networks, yet only mapped the hidden state of one layer to the output module by setting all other layer weights to zero. Outputs from these "masked" language models represented the predictiveness of words at one of five timescales.

As the primary metric of predictiveness, we calculated the degree of surprisal associated with the occurrence of a word given its context (i.e., negative logarithm of the probability assigned to the *actual* next word). The surprisal evoked by an incoming word indexes the amount of information that was not predictable from the context represented at a specific timescale (*35*, *36*). Of note, surprisal was considerably higher at longer timescales in the LSTM ($p < 0.001$, Cohen's $d = 2.43$; compared to slopes drawn from surprisal shuffled across timescales) but remained stable across

172 timescales in the HM-LSTM ($p = 0.955$, $d = 0.05$; direct comparison LSTM vs. HM-LSTM: $p < 0.001$, $d =$
173 2.7; Figure 2A).
174 To determine the temporal integration window of each timescale, we scrambled input to the
175 network at different granularities corresponding to a binary logarithmic increase in the length of
176 intact context (i.e., 1–256 words). The LSTM showed no increase of temporal integration windows at
177 higher layers (LSTM: $p = 0.219$, $d = 0.11$). In contrast, in the HM-LSTM, surprisal decreased more
178 strongly at longer compared to shorter timescales as more intact context became available (HM-
179 LSTM: $p = 0.027$, $d = 0.73$; LSTM vs. HM-LSTM: $p < 0.001$, $d = 0.76$; Figure 2B).
180 Our secondary metrics expressed the predictability of a word in relation to other words. First, the
181 entropy of the probability distribution over candidate words indicates the difficulty to make a
182 definite prediction on an upcoming word. Second, the dissimilarity of vector representations (or
183 embeddings) coding for the constituent features of the predicted and actual next word (Product-
184 moment correlation) indicates conceptual (un-)relatedness.
185 We derived surprisal, entropy and dissimilarity associated with single words in the story from
186 masked models at each of five timescales and from "full" models across all timescales, separately for
187 each language model (see Figure S2 for correlations of metrics). All features were convolved with the
188 hemodynamic response function, and we will collectively refer to them as "features of
189 predictiveness" from here on.

190 **Higher model-derived surprisal of words slows down reading**

191 To test the behavioural relevance of model-based predictiveness, another 26 participants performed
192 a self-paced reading task where they read the transcribed story word-by-word on a noncumulative
193 display and pressed a button as soon as they had finished reading.
194 When regressing response speed onto time-lagged features of predictiveness and a set of
195 nuisance regressors (e.g., word length and frequency), we found that—as expected—reading speed
196 slowed down for words determined as more surprising by language models given the full context
197 across all timescales (Figure 2C).
198 Further, we predicted response speed on held-out testing data and z-scored the resulting
199 encoding accuracy (i.e., Product-moment correlation of predicted and actual response speed) to a
200 null distribution drawn from scrambled features of predictiveness while only keeping nuisance
201 regressors intact. This yielded the unique contribution of the predictiveness of words (i.e., surprisal,
202 entropy and dissimilarity) to reading speed, which was significant for both language models (LSTM:
203 $p < 0.001$, $d = 1.51$; HM-LSTM: $p < 0.001$, $d = 1.64$, LSTM vs. HM-LSTM: $p = 0.975$, $d = 0.35$). Together,
204 these findings suggest that both language models picked up on processes of speech prediction that
205 shape behaviour.

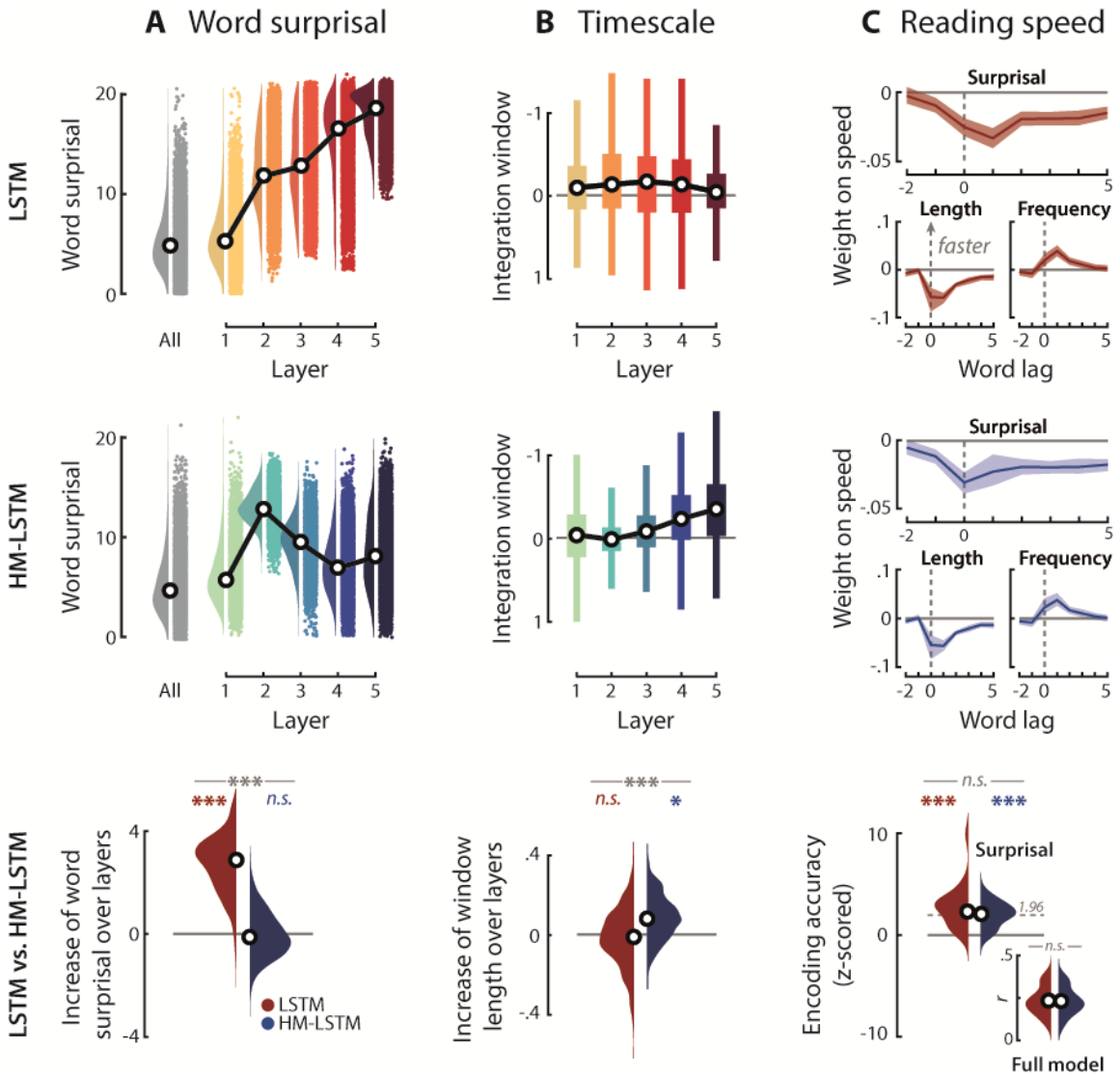

Figure 2. Evaluating model-derived surprisal. (A) Top: Word surprisal derived from the "full" LSTM model including all layers (grey distribution) and from single layers of "masked" LSTM models (coloured distributions); black circles represent grand-median surprisal. Middle: Same as above, but for the sparsely updating HM-LSTM. Bottom: We fit linear functions to word surprisal across layers and compared resulting slope parameters to null distributions drawn from shuffled layers (LSTM: red, HM-LSTM: blue) and between language models (LSTM vs. HM-LSTM: level of significance shown in grey). **(B)** Top: Input to the LSTM was scrambled at different granularities corresponding to an increase in the length of intact context (i.e., 1–256 words). For each layer of the LSTM, linear functions were fit to word surprisal across these context windows. A negative slope parameter indicates a stronger benefit (or lower surprisal) from longer context (i.e., larger integration window or timescale). Middle: Same as above, but for the sparsely updating HM-LSTM. Bottom: Linear fit to integration windows across layers, indicating the benefit of higher layers from longer context. **(C)** Top: Speed in a self-paced reading task was modelled as a function of time-lagged predictiveness and a set of nuisance regressors (i.e., length, frequency and number of words as well as content vs. function words). The upper weight profile shows the temporal dynamics of the surprisal effect in the full LSTM model. To illustrate the magnitude of the surprisal effect, we also show weight profiles for two factors known to have a major impact on reading speed (*49*), that is word length (bottom, left) and word frequency (bottom, right); positive weights indicate an increase in response speed; error bands represent ±SEM. Middle: Same as above, but for the sparsely updating HM-LSTM. Bottom: Encoding accuracy in the self-paced reading task uniquely explained by the predictiveness of context (standardized to scrambled features of predictiveness); dotted grey line indicates critical significance level for single participants. Inset shows non-standardized encoding accuracies. *** $p < 0.001$, * $p < 0.05$, *n.s.*: not significant.

227 **Selecting temporo-parietal regions of interest involved in speech processing**

228 We hypothesized that the speech prediction hierarchy is represented as a gradient along the
229 temporo-parietal pathway. This rather coarse region of interest was further refined to only include
230 regions implicated in processing of the listening task.

231 To this aim, we calculated pairwise intersubject correlations (*50*), which revealed consistent
232 cortical activity across participants in a broad bilateral language network. Responses were most
233 prominently shared in auditory association cortex and lateral temporal cortex as well as premotor
234 cortex, paracentral lobule and mid cingulate cortex (Figure 3A). Crucially, as sound textures
235 presented in the competing stream were randomly ordered across participants, this approach
236 allowed us to extract shared responses specific to the speech stream.

237 The context representations necessary to make predictions are thought to be located in temporal
238 and parietal cortex regions, while most other regions showing increased intersubject correlations
239 are thought to be implicated in networks of cognitive control and action (*51*). Therefore, we limited
240 all further analyses to those temporal and parietal parcels (*52*) in which more than 80 % of vertices
241 yielded a significant intersubject correlation. We tested for significance by ranking the median
242 group-level intersubject correlation of single vertices against a bootstrapped null distribution (p_{FDR}
243 < 0.01). The cortical sheet of the six parcels determined as regions of interest (ROI) was flattened,
244 resulting in a two-dimensional plane spanned by an anterior-posterior and inferior-superior axis
245 (Figure 3B).

246 We expected gradients of speech prediction to unfold along the inferior-superior axis, that is,
247 from temporal to parietal areas.

248 **Differential tuning to continuously versus sparsely updated timescales of surprisal in**
249 **temporo-parietal cortex**

250 In an encoding analysis, we regressed hemodynamic responses of single vertices in the temporo-
251 parietal ROI onto model-based features of predictiveness (i.e., surprisal, entropy and dissimilarity at
252 single timescales and for full models) as well as a set of acoustic and linguistic nuisance regressors.
253 For each language model, we extracted five temporo-parietal maps showing the timescale-specific
254 regression weights of surprisal.

255 When performing spatial clustering on these weight maps (p_{vertex} and $p_{cluster} < 0.05$; compared to
256 scrambled surprisal by means of a cluster-based permutation test), we found large positive clusters
257 in both hemispheres for surprisal at shorter timescales of the LSTM (Figure 4A, yellow outlines) but,
258 if at all, only focal clusters at longer timescales (Figure 4A, red outlines). Hence, temporo-parietal
259 activity primarily increased in response to words that were less predictable by the context provided
260 at shorter, continuously updated timescales. In contrast, we found clusters of distinct polarity,
261 location and extent for surprisal at all timescales of the HM-LSTM (Figure 4B). This suggests that even

longer timescales had the potency to modulate temporo-parietal activity when they were sparsely updated.

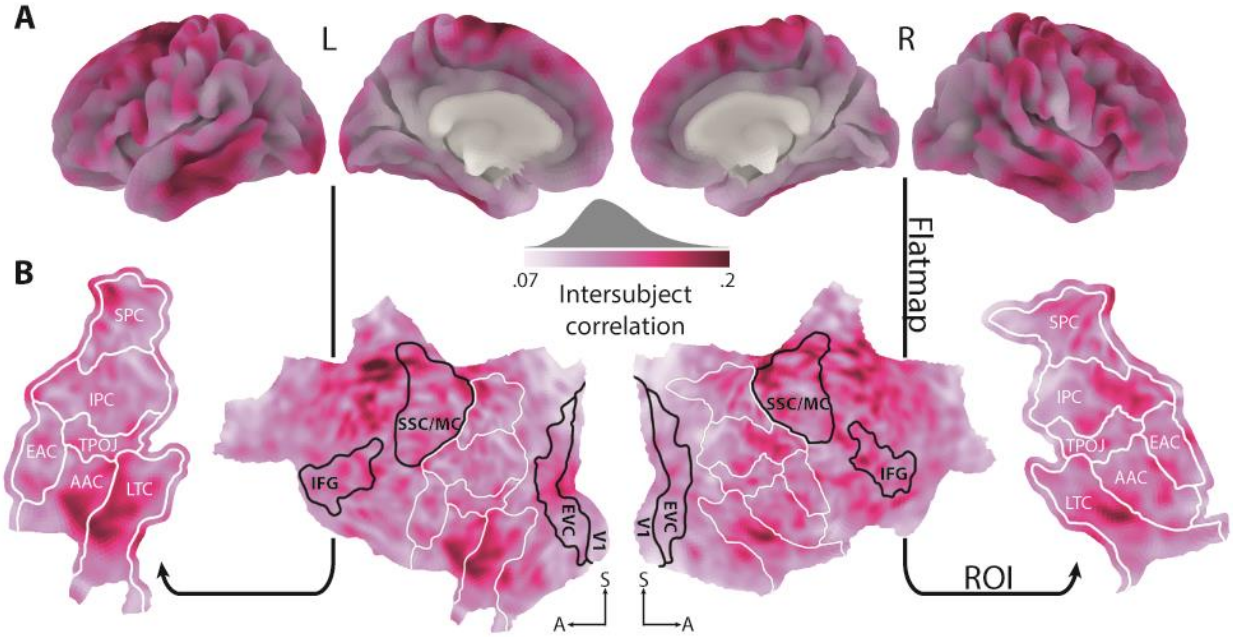

Figure 3. Selection of regions of interest. (A) When listening to a story against background noise, pairwise intersubject correlations showed stronger synchronization of BOLD activity in cortical areas implicated in the language network. **(B)** The cortical surface was flattened. All temporal and parietal parcels (*52*) highlighted by white outlines were included as regions of interest (ROI) in the following analyses. Black outlined parcels serve as reference point only. EAC: early auditory cortex, AAC: auditory association cortex, LTC: lateral temporal cortex, TPOJ: temporo-parieto-occipital junction, IPC: inferior parietal cortex, SPC: superior parietal cortex, V1: primary auditory cortex, EVC: early visual cortex, IFG: inferior frontal gyrus, SSC/MC: somatosensory and motor cortex. Maps were smoothed with an 8 mm FWHM Gaussian kernel for illustration only.

Sparsely updated timescales of surprisal evolve along a temporo-parietal processing hierarchy

To probe the organization of surprisal at different timescales along a temporo-parietal gradient, we collapsed across the anterior-posterior axis of single weight maps and selected the local maximum with the largest positive value on the inferior-superior axis. Fitting a linear function to these peak coordinates of timescales, we found flat slope parameters indicating random ordering of LSTM timescales in both hemispheres (left: $p = 0.458$, $d = -0.15$; right: $p = 0.716$, $d = -0.07$; compared to slopes drawn from coordinates scrambled across timescales, Figure 4C). Conversely, we found steep positive slopes for the HM-LSTM in both hemispheres (left: $p < 0.001$, $d = 0.72$; right: $p < 0.001$, $d = 0.75$; Figure 4C), reflecting the representation of surprisal at longer timescales in more parietal regions. The grand-average slope of the HM-LSTM in the left hemisphere indicates that surprisal at a sparsely updated timescale is represented 12 mm superior to its directly preceding timescale along the unfolded temporo-parietal surface. Most relevant, this finding was underpinned by a significant difference of slope parameters between the LSTM and HM-LSTM (left: $p = 0.005$, $d = 0.9$; right: $p <$

286 0.001, $d = 0.89$; Figure 4C), demonstrating a temporo-parietal processing hierarchy of word surprisal
287 that preferably operates on sparsely updated timescales.

288 The absence of a gradient for surprisal at continuously updated timescales was corroborated
289 when specifically targeting the dorsal processing stream. To this aim, we confined surprisal at the
290 first timescale to peak in temporal regions and surprisal at all other timescales to peak superior to
291 the first timescale in more parietal regions. Slope effects along the dorsal stream largely complied
292 with those found in the unconstrained approach (LSTM: all $p \geq 0.134$; HM-LSTM: all $p \leq 0.006$; LSTM
293 vs. HM-LSTM: all $p \geq 0.103$), thereby ruling out the possibility that the presence of a competing
294 ventral stream obscured the consistent ordering of timescales.

295 Additionally, rotating weight maps around the full circle in steps of 5° before collapsing across
296 the first dimension showed that surprisal at sparsely updated timescales was processed along a
297 broad gradient, roughly covering rotations from 0° to -90° in both hemispheres (p_{vertex} and $p_{cluster}$ <
298 0.05; compared to scrambled surprisal by means of a cluster-based permutation test). No effect was
299 found for the LSTM.

300 In a complementary decoding approach (see Supplementary Text S7), we reconstructed surprisal
301 at the five timescales from patterns of neural activity in single regions of interest (i.e., temporo-
302 parietal parcels). Contrasting decoding accuracies between language models, left-hemispheric early
303 auditory cortex contained more information on surprisal at medium, sparsely updated timescales,
304 whereas inferior and superior parietal cortex preferentially represented surprisal at long, sparsely
305 updated timescales (Figure S3). This finding converges with the organization of the gradient
306 described earlier for sparsely updated but not continuously updated timescales of surprisal.

307 Unlike for the sparsely updated timescales of surprisal, neither the timescales of entropy (all $p \geq$
308 0.583) nor dissimilarity (all $p \geq 0.623$) organized along a dorsal gradient (Figure S4 and S5). Further,
309 effects of HM-LSTM timescale surprisal were dissociable from a simple measure of semantic
310 incongruence between words in the story and their preceding context at five timescales
311 logarithmically increasing in length (all $p \geq 0.5$; Product-moment correlation of target and average
312 context embedding). This highlights the specificity of the observed gradient to prediction processes
313 in general and word surprisal in particular.

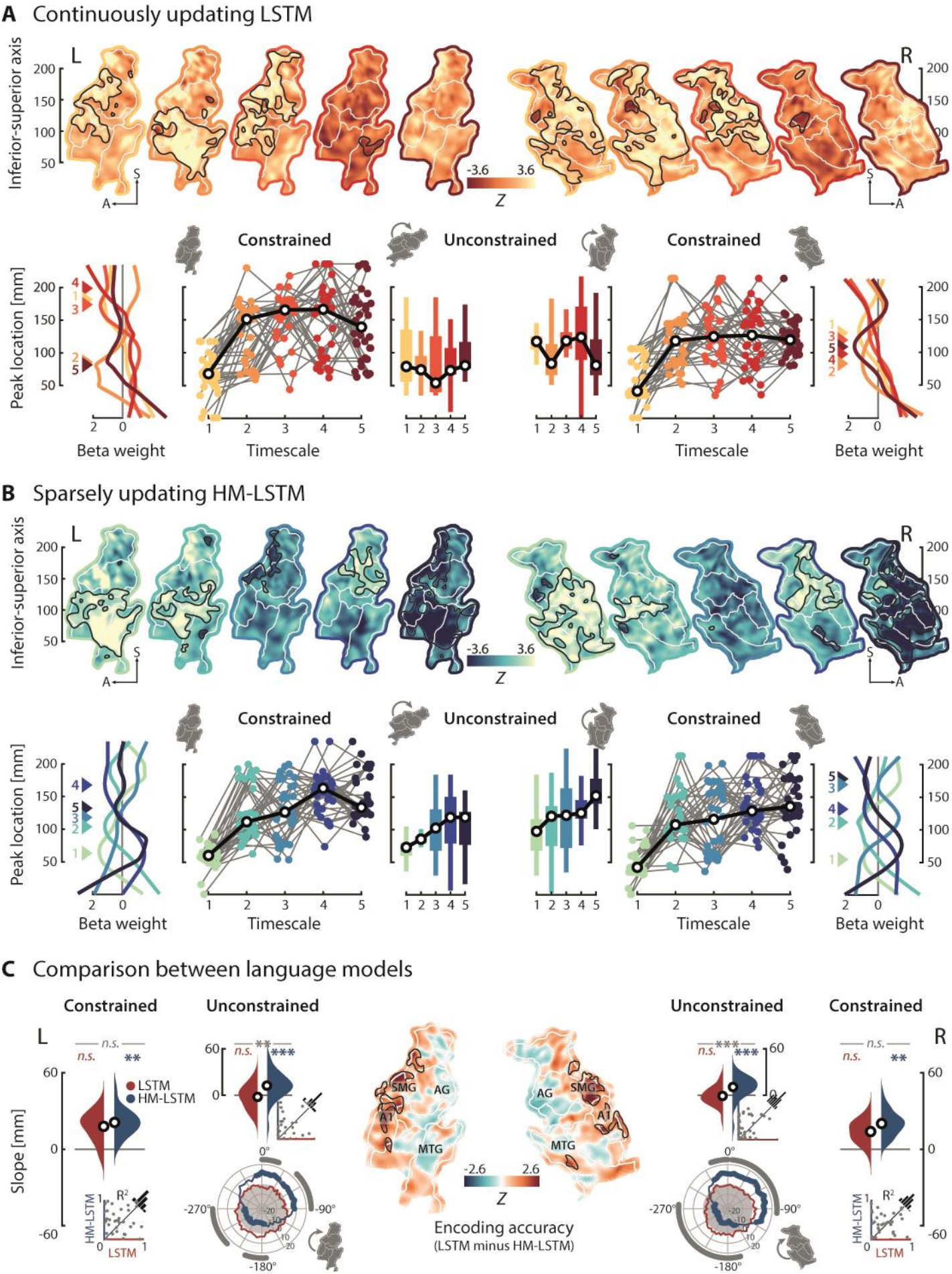
A Continuously updating LSTM
Inferior-superior axis
L
R
-3.6
3.6
Z
Constrained
Unconstrained
Constrained
Peak location [mm]
Beta weight
Timescale
B Sparsely updating HM-LSTM
C Comparison between language models
Slope [mm]
LSTM
HM-LSTM
n.s.
SMG
AG
A1
MTG
-2.6
2.6
Encoding accuracy
(LSTM minus HM-LSTM)

Figure 4. Encoding surprisal at multiple timescales. (A) Temporo-parietal BOLD time series were mapped onto the predictiveness of speech derived at five timescales from the continuously updating LSTM. Top row: Maps show *z*-values from timescale-specific weights of surprisal tested against zero; positive *z*-values indicate an increase of BOLD activity in response to more surprising words, whereas negative *z*-values indicate a decrease of BOLD activity. Black outlines indicate significant clusters (compared to a null distribution drawn from scrambled surprisal); white outlines indicate parcels; coloured outlines indicate short (light yellow) to long (dark red) timescales, separately for the left (L) and right hemisphere (R). Bottom row: For surprisal at each timescale, we determined its peak coordinate along the inferior-superior axis (coloured triangles numbered according to timescale), here shown for grand-average weight profiles for illustration only. Testing for a processing hierarchy along the dorsal stream, timescales were constrained to peak superior to the first timescale; coloured dots connected by grey lines represent peak coordinates of single participants; black circles represent grand-median peak coordinates. Testing for a processing hierarchy in the unconstrained approach, timescales were allowed to peak at any location. Maps were rotated around the inferior-superior axis (here shown for -45°) to test for the spatial specificity of the effect. **(B)** Encoding maps and timescale-specific peak locations for the sparsely updating HM-LSTM. **(C)** Linear functions were fit to peak coordinates across timescales and resulting slope parameters were compared to null distributions drawn from scrambled coordinates (LSTM: red, HM-LSTM: blue) and between language models (LSTM vs. HM-LSTM: level of significance shown in grey), separately for the constrained and unconstrained (unrotated) approach in both hemispheres. Black circles represent grand-average slope parameters; insets depict coefficients of determination for linear fits of single participants. Additionally, we tested for slope effects around the full circle (rose plots); grey areas indicate negative, white areas positive slope parameters; fat coloured lines indicate significant slope clusters of single language models; fat grey lines indicate significant differences of the slope between language models. Temporo-parietal encoding accuracies displayed on ROI maps were z-scored to null distributions drawn from scrambled features of predictiveness and compared between language models; black outlines indicate significant clusters; maps were smoothed with an 8 mm FWHM Gaussian kernel for illustration only; SMG: supramarginal gyrus, AG: angular gyrus, A1: primary auditory cortex; MTG: middle temporal gyrus. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, *n.s.*: not significant.

To determine the contribution of predictiveness to overall encoding accuracy on held-out data, we z-scored accuracies relative to null distributions drawn from scrambled features of predictiveness while keeping additional (spectro-temporal) acoustic and linguistic nuisance regressors intact (for encoding accuracies of these other regressors see Figure S6). Interestingly, the LSTM produced—in comparison to the HM-LSTM—better predictions in early auditory cortex and supramarginal gyrus (p_{vertex} and $p_{cluster} < 0.05$; cluster-based permutation paired-sample *t*-test; Figure 4C). On the other hand, predictions of the HM-LSTM seemed slightly more accurate than for the LSTM along middle temporal gyrus, temporo-parieto-occipital junction and angular gyrus, even though not statistically significant. Taking into account the broad clusters found earlier specifically for surprisal at shorter (but not longer) LSTM timescales, this underscores the notion that surprisal at continuously updated timescales takes full effect only in earlier processing stages, whereas the sparsely updating processing hierarchy evolves to more posterior parietal regions.

Surprisal at sparsely updated timescales gates connectivity along the processing hierarchy

After establishing the temporo-parietal processing hierarchy, we examined the modulatory effect of surprisal on connectivity between peak locations of timescales taken from the encoding analysis. Following the assumptions of predictive coding (*16*), the bottom-up information flow from a brain area representative of a shorter timescale to its immediately neighbouring, longer timescale should

increase when the prediction error at the shorter timescale is higher, thereby initiating updates to predictions at the longer timescale. In turn, higher timescale-specific surprisal should increase the top-down information flow to the immediately shorter timescale, thereby indicating the backward-pass of updated predictions. Such a surprisal-gated hierarchy is thought to allow for efficient mapping between timescales.

To this aim, we created psychophysiological interactions (*53*) between the BOLD response at the peak location of one timescale and word surprisal at the same timescale. The BOLD response at the peak location of each (target) timescale was regressed onto psychophysiological interactions of all other (predictor) timescales (Figure 6A). A positive weight indicates increased coupling between two peak locations when surprisal at the predictor timescale is high.

We hypothesized that coupling between brain regions representing surprisal at two neighbouring timescales increases when one timescale becomes unpredictive. Numerically, this can be expressed by setting the weights of neighbouring timescales to 1 and all other predictor weights to -1 (Figure 6B). This hypothesized pattern of weights was not matched by the weights observed for the LSTM (left: $p = 0.83$, $d = 0.27$; right: $p = 0.348$, $d = 0.1$; Euclidean distance compared to null distributions drawn from target BOLD activity shifted in time), which was expected given that surprisal at the continuously updated timescales was not organized along a gradient in the first place. Critically, for surprisal at sparsely updated timescales of the HM-LSTM, surprisal-modulated connectivity in the left hemisphere not only matched our hypothesis (left: $p = 0.032$, $d = 0.5$; right: $p = 0.853$, $d = 0.23$) but also matched our hypothesis better than the LSTM (left: $p = 0.001$, $d = 0.79$; right: $p = 0.902$, $d = 0.3$).

To specify the overall directionality of information flow, we separately averaged weights of top-down modulations (i.e., predictor weights of surprisal at timescales longer than respective target timescales) and bottom-up modulations. We found no difference between the modulatory strength of these top-down and bottom-up connections (LSTM left: $p = 0.184$, $d = 0.23$; right: $p = 0.839$, $d = 0.4$; HM-LSTM left: $p = 0.408$, $d = 0.4$; right: $p = 0.367$, $d = 0.16$).

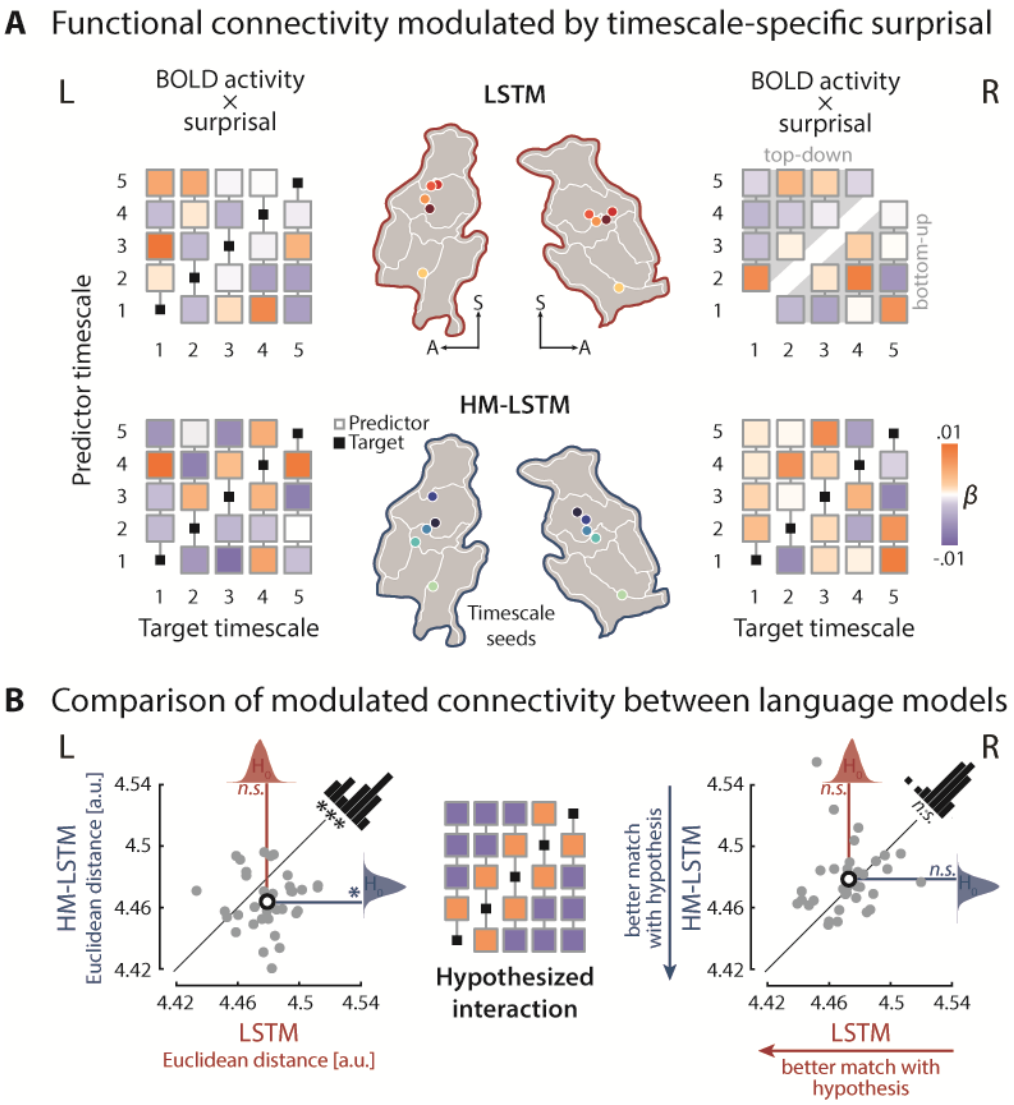

Figure 5. Surprisal-dependent modulation of effective connectivity. (A) A sphere of 6 mm was centred on median peak locations of the timescales of surprisal as defined in the encoding analysis (coloured circles on temporo-parietal maps) and BOLD responses were averaged within these timescale seeds. BOLD time series at one (target) timescale were regressed onto psychophysiological interactions of all other (predictor) timescales (i.e., pointwise product of timescale-specific BOLD and surprisal time series). For each target seed, we added a column vector of timescale-specific predictor weights to a 5-by-5 matrix with an empty main diagonal. Matrices were created separately for each language model (top: LSTM, bottom: HM-LSTM) and hemisphere. The upper triangle of a matrix indicates top-down, the lower triangle bottom-up information flow. **(B)** A hypothesized matrix of psychophysiological interactions was created, with positive weights on diagonals below and above the main diagonal (orange squares), indicating increased connectivity between surprisal at neighbouring timescales when surprisal is high. The Euclidean distance between observed and hypothesized matrices was compared to null distributions of distances drawn from target timescales shifted in time (LSTM: red density plot, HM-LSTM: blue) and between language models (LSTM vs. HM-LSTM: level of significance shown in black), separately for the left and right hemisphere; grey dots indicate distances of single participants; black circles indicate mean distances. * $p < 0.05$; *** $p < 0.001$; *n.s.*: not significant.

399 # Discussion

400 How are the complex temporal dependencies underlying natural speech processed in the brain to
401 inform predictions of upcoming speech? In the present study, we simulated these prediction
402 processes in two language models (LSTM vs. HM-LSTM), which critically differed in how often
403 semantic context representations are updated at multiple, hierarchically organized timescales.

404 Surprisal as derived from both language models modulated reading times in the behavioural
405 reading task to a similar degree. However, hemodynamic brain responses to surprisal during the
406 listening task differed between models: In line with our initial hypothesis, temporo-parietal regions
407 hierarchically encoded the (sparsely updated) event-based surprisal provided by the layers of the
408 HM-LSTM, with surprisal at longer timescales represented in inferior parietal regions. Moreover,
409 higher timescale-specific surprisal based on the HM-LSTM increased connectivity from receptive
410 windows of surprisal at a given timescale to their immediately neighbouring (shorter or longer)
411 timescales.

412 Together, these results provide evidence for the neurobiological parsimony of an event-based
413 processing hierarchy. In the present data, this was expressed in the simultaneous neural
414 representation of surprisal at multiple timescales, and in surprisal dynamically gating the
415 connectivity between these timescale-specific receptive windows.

416 **The event-based organization of context as a foundation for language prediction**

417 The spatial organization of timescale-specific receptive windows of surprisal observed in the present
418 study converges with previous results, where bilateral primary auditory cortex coded for relatively
419 shorter timescales (e.g., words) and inferior parietal cortex coded for longer timescales (e.g.,
420 paragraphs; Ref. (*27*)). Critically, this spatial overlap was found despite targeting different aspects of
421 speech processing.

422 In our study, neural responses were expressed as a function of timescale-specific word surprisal,
423 a proxy tapping into prediction processes. In other studies, receptive windows were based on the
424 (in-)consistency of neural activity across participants in response to speech input at varying
425 timescales (*54*), which is typically linked to working memory formation. This implies that the same
426 neural system most likely fulfils distinct functions: Temporal receptive windows have been
427 suggested to store timescale-specific context in working memory, and, in parallel, exploit this
428 context to process information in the present (*9*). In line with this theoretical account, our results
429 suggest that timescale-specific memory representations serve as the basis for the generative models
430 shaping predictions of upcoming speech.

431 The here observed temporo-parietal gradient of surprisal at sparsely updated representations of
432 context is specifically well in line with accounts of neural event segmentation (*55*, *56*), and with the

433 notion of hierarchical multiscale network architectures more generally (here, HM-LSTM). Taking a
434 sentence from our listening task as an example, "The wild wine was called ink grapes" represents a
435 brief event where the narrator describes how the bluish black of the grapes in a backyard reminded
436 her of the color of the night. At the same time, the sentence is preceded by a larger event of the
437 author walking the streets on the way to her parents' house. Both events can be combined into an
438 even larger event of the author visiting her Romanian hometown.

439 Of note, the events modelled by the HM-LSTM do not necessarily concur with human event
440 annotations and are not fully independent of previous events due to a leaky network architecture.
441 Notwithstanding these limitations, we can draw an analogy between the HM-LSTM architecture and
442 neural event segmentation in two decisive points. First, the boundary detector allows revealing an
443 event structure of context, similar to an increase of neural activity indexing prediction errors at event
444 boundaries (*57*, *58*). Second, the sparse updates to higher processing stages at event boundaries
445 allow retaining multiple, stable context representations in memory, similar to temporo-parieto-
446 occipital receptive windows reflecting the hierarchical event structure during movie watching (*41*).
447 We directly tie in with this result by showing that such hierarchical, event-based context enables
448 neural prediction processes.

449 What are the computational and mechanistic implications of this contextual architecture for the
450 prediction of speech? Somewhat paradoxically, event models have been referred to as "an added
451 burden for an organism" (*59*). This argument is certainly plausible with regard to the size of the
452 parameter space, which increases in an artificial or, likewise, biological neural network by
453 introducing an additional boundary detector. At the expense of model parsimony, however, such an
454 event-based network allows for fewer updates in comparison to continuously updating networks
455 like the LSTM, where each new input to the model elicits computationally complex updates to all
456 timescales.

457 The trade-off between computational costs of boundary detector versus update frequency is
458 nicely illustrated by the fact that the sparsely updating HM-LSTM is considerably faster in making
459 predictions than the continuously updating LSTM (*42*). Thus, from a functional perspective, keeping
460 layered representations of multiple events in memory allows to efficiently draw on diverse
461 information to make predictions on upcoming speech.

462 **Context-dependent surprisal as a gating mechanism for predictions and prediction errors**

463 The hallmark of prediction processes in our data is the increase in reading times and neural activity
464 observed in response to more surprising input (*21*, *60*). There are different computational ways in
465 which this "expectation suppression" (*61*) can be realized, namely integration difficulty, neural
466 sharpening, and predictive coding.

467 One take on expectation suppression is that surprising sensory input is more difficult to integrate
468 into already existing representations of context because it conveys a relatively larger amount of new
469 information (*62*). The architecture of the sparsely updating HM-LSTM dictates that new information
470 is integrated into timescale-specific representations only at event boundaries, thereby suggesting
471 that integration difficulty should arise primarily on an event-by-event basis. As surprisal indeed
472 varies on a word-by-word basis, it is less likely that integration difficulty accounts for our effects of
473 event-based surprisal. However, accounts of integration difficulty and predictive processing are not
474 mutually exclusive. The other two accounts both assume that expectation suppression is indicative
475 of prediction processes but differ in how these processes are thought to be implemented in the
476 brain.

477 Sharpening accounts argue that *un*expected components of sensory input are suppressed via
478 feedback predictions (*63*), while accounts of predictive coding (*16*, *64*) suggest that the brain filters
479 out (or "dampens") expected components of sensory input. Both accounts assume that predictive
480 processes lead to an overall decrease of neural activity to incoming stimuli. The similarity between
481 hypothesized response patterns makes it notoriously hard to disentangle those accounts (*32*).

482 Notably, however, a distinguishing feature of predictive coding is the specificity of feedforward
483 prediction error signals, which can be captured by modelling effective connectivity between
484 receptive windows of timescale-specific surprisal. In agreement with the hierarchical information
485 flow laid out in predictive coding (*16*), surprisal in our study modulated connectivity via bidirectional
486 links between neighbouring receptive windows of longer and shorter event-based timescales in the
487 left hemisphere (Figure 6).

488 Surprisal in the event-based artificial neural network was modelled as the amount of information
489 an input word conveys that cannot be explained away by the context (or generative model)
490 represented at a specific timescale. Therefore, the increase of feedforward connectivity in response
491 to higher surprisal precisely aligns with the concept of prediction errors in predictive coding (*64*).

492 In addition, the increase of feedback connectivity in response to higher surprisal accords with an
493 electrocorticography study in macaques by Chao and colleagues (*24*). The study showed that
494 prediction errors evoked in tone sequences trigger feedback signals from prefrontal to anterior
495 temporal and early auditory cortex in alpha and beta frequency bands. Extending these previous
496 results, our findings suggest that surprisal initiates bottom-up prediction errors, indicative of
497 imprecise predictions, and top-down updates to predictions at processing stages of shorter events
498 to facilitate perception of new words.

499 As an interim conclusion, our findings have two important implications for frameworks of
500 prediction and prediction error: First, we show that a multi-layered hierarchy of predictive coding
501 (e.g., Ref. (*18*)) applies well to temporo-parietal language processing. Second, predictive coding

502 remains a viable account of neural processing, also when put to test using complex temporal
503 dependencies underlying real-life stimuli.

504 **Implications for a larger network perspective on the event-based prediction hierarchy**

505 Dual stream models of language propose that speech processing is organized along a ventral and a
506 dorsal stream (*65*). In the present study, we found a hierarchy of speech prediction along the dorsal
507 stream, which emanated from early auditory cortex and extended well into parietal cortex (Figure
508 4B).

509 This result may seem at odds with other studies showing an additional mirror-symmetric ventral
510 gradient, in which more complex speech features are represented in more anterior temporal regions
511 (*66*). The ventral stream has been proposed to chunk speech features into increasingly abstract
512 concepts irrespective of their temporal presentation order (*67*). In contrast, we here modelled
513 context representations by respecting the temporal order of words, that is, the HM-LSTM integrates
514 incoming words into an event until a change in situational features of words indicates the end of an
515 event and a new event is created. Hence, the ventral stream may contribute to hierarchical speech
516 prediction by exploiting another, more nested facet of context.

517 The inferior frontal gyrus (IFG), alongside premotor cortex, is deemed the apex of the dorsal
518 stream (*65*), yet we here considered only the role of temporo-parietal cortex in speech prediction.
519 Previous studies showed that activity in IFG relies on longer timescales of speech being intact (*27*,
520 *68*), that connectivity between IFG and superior temporal gyrus is driven by expectations (*69*, *70*),
521 and that right IFG is sensitive to the violation of non-local regularities (*71*, *72*). While this suggests an
522 interplay between frontal and temporo-parietal regions in hierarchical speech prediction, the
523 precise anticipatory mechanisms IFG exerts cognitive control over are just as unclear as how top-
524 down cognitive control and bottom-up sensory input are balanced along the hierarchy.

525 Another open question is the cross-linguistic generalizability of the observed speech prediction
526 hierarchy. The present study focused on German, which is—like any other language—characterized
527 by its unique linguistic structure shaping also neural computations. For example, alphabetic
528 languages like German versus tonal languages like Chinese engage different neural systems for
529 phonological processing (*73*). On the other hand, higher-level linguistic processing in temporal and
530 parietal cortex has been shown to be consistent across a diverse set of languages and language
531 families (*74*). As the event structure of speech relies on such higher-level semantic representations,
532 this opens up the possibility that event-based predictive processing along the temporo-parietal
533 pathway is a universal function of language comprehension.

534 Beyond short-term semantic context, also long-term knowledge facilitates speech prediction. In
535 theory, both memory systems can be couched into the larger framework of the dual reference frame
536 system (*75*), where flexible sensory knowledge in parietal cortex interacts with stable conceptual

knowledge in hippocampus. Consistent with the key characteristics of the speech prediction hierarchy, hippocampus codes for boundaries in the environment (*76*), hierarchically organizes memories (*77*) and engages in predictive coding (*78*). As parietal cortex has been shown to interface with hippocampus at event boundaries of longer timescales during movie watching (*41*), we speculate that the hierarchy of speech prediction might extend from receptive windows in parietal cortex to hippocampus.

Importantly, the event-based prediction hierarchy relies on a set of neural computations—i.e., event segmentation, temporal receptive windows, predictive coding—available beyond the domain of language. For example, taking a walk through the city and then taking a break in the garden constitutes two events that are not just distinguishable by the change of location when presented as a narrated story but also by the change of visual (sight of traffic lights vs. trees) or olfactory features (smell of car exhausts vs. flowers) when experienced in a natural environment. Our results thereby encourage future studies to probe the generalizability of the event-based prediction hierarchy to other species, sensory modalities, and cognitive functions.

The present study bridges the gap between the hierarchical, temporally structured organization of context in language comprehension on the one hand and the more general principles of hierarchical predictive processing in cerebral cortex on the other hand.

Combining continuously narrated speech, artificial neural networks, and functional MRI building on these networks' output allowed us, first, to sample the natural dynamic range of word-to-word changes in predictiveness over a multi-level hierarchy. Second, we were able to systematically compare the neural effects of different contextual updating mechanisms.

Our data demonstrate that the prediction processes in language comprehension build on an event-based organization of semantic context along the temporo-parietal pathway. Not least, we posit that such an event-based organization provides a blueprint for a contextually diverse, yet computationally efficient network architecture of anticipatory processing in complex naturalistic environments.

Materials and Methods

Participants

Thirty-seven healthy, young students took part in the fMRI listening study. The final sample included *N* = 34 participants (18–32 years; *M* = 24.65; 18 female), as data from one participant was excluded from all analyses due to strong head movement throughout the recording (mean framewise displacement > 2 SD above group average (*79*)) and two experimental sessions were aborted because participants reported to not understand speech against noise. Another 26 students (19–32 years; *M* = 23.54; 17 female) took part in the behavioural self-paced reading study.

All participants were right-handed German native speakers who reported no neurological, psychiatric or hearing disorders. Participants gave written informed consent and received an expense allowance of €10 per hour of testing. The study was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committee of the University of Lübeck.

Stimulus materials

As a speech stimulus in the fMRI listening study, we used the first 64 minutes of an audio recording featuring Herta Müller, a Nobel laureate in Literature, reminiscing about her childhood as part of the German-speaking minority in the Romanian Banat ("Die Nacht ist aus Tinte gemacht", 2009).

To emulate an acoustically challenging scenario in which listeners are likely to make use of the predictability of speech (*80*), this recording was energetically masked by a stream of concatenated five-second sound textures at a signal-to-noise ratio of 0 dB. Sound textures were synthesized from the spectro-temporal modulation content of 192 natural sounds (i.e., human and animal vocalizations, music, tools, nature scenes (*81*)), so that the noise stream did not provide any semantic content potentially interfering with the prediction of upcoming speech. The order in which sound textures were arranged was randomized across participants. For more details on how sound textures in the present experiment were generated and how they were processed in auditory cortex, see Ref. (*48*).

The monaural speech and noise streams were sampled to 44.1 kHz and custom filters specific to the left and right channel of the earphones used in the fMRI experiment were applied for frequency response equalization. Finally, speech-in-noise stimuli were divided into eight excerpts à 8 minutes, which served as independent runs in the experiment.

A trained human external agent literally transcribed the speech stream. The text transcript comprised 9,446 words, which were used as stimuli in the self-paced reading task and as input to our language models. To automatically determine onset and offset times of all spoken words and phonemes, we used the web service of the Bavarian Archive for Speech Signals (*82*): First, the text transcript was transformed to a canonical phonetic transcript encoded in SAM-PA by the G2P

module. Second, the most likely pronunciation for the phonetic transcript was determined by a Markov model and aligned to the speech recording by the MAUS module. Fourteen part-of-speech tags were assigned to the words in the text transcript using the pre-trained German language model de_core_news_sm (2.2.5) from spaCy (https://spacy.io/). Based on these tags, words were classified as content or function words. Word frequencies were derived from the subtitle-based SUBTLEX-DE corpus (*83*) and transformed to standardized Zipf values (*84*) operating on a logarithmic scale from about 1 (word with a frequency of 1 per 100 million words) to 7 (1 per 1,000 words). The Zipf value of a word not observed in the corpus was 1.59 (i.e., smallest possible value).

Experimental procedures

Behavioural self-paced reading task. While the transcribed story was presented word-by-word on a noncumulative display, participants had the task to read each word once at a comfortable pace and quickly press a button to reveal the next word as soon as they had finished reading. A timeout of 6 seconds was implemented. The time interval between word appearance and button press was logged as the reading time. After each run, participants answered three four-option multiple-choice questions on the plot of the story (performance: *Ra* = 58.33–100 % correct, *M* = 79.17 %, *SD* = 10.87 %) and took a self-paced break. In total, each participant completed four out of eight runs, which were randomly selected and presented in chronological order. Throughout the reading task, we recorded movement and pupil dilation of participants' left eye at a sampling rate of 250 Hz in one continuous shot with an eye tracker (EyeLink 1000, SR Research). Eye tracking data were not analysed in the present study.

The experiment was controlled via the Psychophysics Toolbox (*85*) in MATLAB (R2017b, MathWorks). All words were presented 20 % off from the left edge of the screen in white Arial font on a grey background with a visual angle of approximately 18°. Participants used a response pad (URP48, The Black Box ToolKit) to navigate the experiment with their right index finger. The experimental session took approximately 40 minutes.

Functional MRI listening task. We instructed participants to carefully listen to the story while ignoring the competing stream of sound textures as well as the MRI scanner noise in the background. Each of the eight runs was initialized by 10 baseline MRI volumes after which a white fixation cross appeared in the middle of a grey screen and playback of the 8-minute audio recording started. MRI recording stopped with the end of playback and participants successively answered the same questions used in the self-paced reading task via a response pad with four buttons (HHSC-2x4-C, Current Designs). On average, participants answered 65.5 % of the questions correctly (*Ra* = 38–100 %, *SD* = 15.9 %). There was a 20-second break between consecutive runs.

The experiment was run in MATLAB (R2016b) using the Psychophysics Toolbox. Stimuli were presented at a subjectively comfortable sound pressure level via insert earphones (S14, SENSIMETRICS) covered with circumaural air cushions. The experimenters monitored whether participants kept their eyes open throughout the experiment via an eye tracker.

MRI data acquisition. MRI data were collected on a 3 Tesla Siemens MAGNETOM Skyra scanner using a 64-channel head coil. During the listening task, continuous whole-brain fMRI data were acquired in eight separate runs using an echo-planar imaging (EPI) sequence (repetition time (TR) = 947 ms, echo time (TE) = 28 ms, flip angle = 60°, voxel size = 2.5 × 2.5 × 2.5 mm, slice thickness = 2.5 mm, matrix size = 80 × 80, field of view = 200 × 200 mm, simultaneous multi-slice factor = 4). Fifty-two axial slices were scanned in interleaved order. For each run, 519 volumes were recorded.

Before each second run, field maps were acquired with a gradient echo (GRE) sequence (TR = 610 ms, TE_1 = 4.92 ms, TE_2 = 7.38 ms, flip angle = 60°, voxel size = 2.5 × 2.5 × 2.75 mm, matrix size = 80 × 80, axial slice number = 62, slice thickness = 2.5 mm, slice gap = 10 %).

In the end of an experimental session, anatomical images were acquired using a T1-weighted (T1w) MP-RAGE sequence (TR = 2,400 ms, TE = 3.16 ms, flip angle = 8°, voxel size = 1 × 1 × 1 mm, matrix size = 256 × 256, sagittal slice number = 176) and a T2-weighted (T2w) SPACE sequence (TR = 3,200 ms, TE = 449 ms, flip angle = 120°, voxel size = 1 × 1 × 1 mm, matrix size = 256 × 256, sagittal slice number = 176).

Modelling the predictiveness of context at multiple timescales

We trained two versions of LSTM networks with five layers to predict the next word in the story given a sequence of semantic context: a continuously updating LSTM where information is fed to a higher layer with each upcoming word, and a competing sparsely updating HM-LSTM where information is fed to a higher layer only at the end of an event. The predictiveness of context at multiple timescales was read out from single layers of both language models for each word in the story presented to participants in experiments. Ultimately, we tested how closely these derivatives of different network architectures match signatures of behavioural and neural prediction processes.

Representing words in vector space. In natural language processing, it is common to represent a word by its linguistic features in the form of high-dimensional vectors (or embeddings). As the German language is morphologically rich and flexibly combines words into new compounds, there are many rare words for which language models cannot learn good (if any) vector representations on the word level. Therefore, we mapped all texts used for training, validating and testing our language models to pre-trained *sub*word vectors publicly available in the BPEmb collection (*86*). These embeddings allow for the representation of *any* word by a combination of 100-dimensional

subwords from a finite vocabulary of 100,000 subwords. We further reduced this vocabulary to those subwords that appeared at least once in any of our texts (i.e., number of subwords in vocabulary $v = 91{,}645$). See Supplementary Text S1 for a detailed description of the BPEmb vocabulary.

Matching our texts to subwords and their respective embeddings in the BPEmb vocabulary yielded the embedded text $t \in R^{w \times e}$ where w is the number of words and $e = 100$ is the number of vector dimensions. On average, a word in the story was represented by 1.07 subwords (*Ra* = 1–6, *SD* = 0.33). As single words were encoded by only one subword in 94.25 % of cases, we will refer to subwords as words from here on.

Architecture of language models. When listening to a story, a fused representation of all spoken words $\{w_1, w_2, \dots, w_p\}$ is maintained in memory and used as context information to make a prediction about the upcoming word w_{p+1}. In natural language processing, this memory formation is implemented via recurrent connections between the states of adjacent neural network cells. The hidden state h_{p-1} stores all relevant context and is sequentially passed to the next cell where it is updated with information from word w_p.

As such a simple recurrent neural network (RNN) tends to memorize only the most recent past, the more complex LSTM (*45*) became a standard model in time series forecasting. In an LSTM cell, the state is split in two vectors: The cell state c_p acts as long-term memory, whereas the hidden state h_p incorporates information relevant to the cell output (i.e., the prediction of the next word). The integration of new information and the information flow between the two memory systems is controlled by three gating mechanisms.

When stacking multiple LSTM cells on top of each other, semantic context gets hierarchically organized in the model, with lower layers coding for short-term dependencies and higher layers coding for long-term dependencies between words. The bottom-up input to the first layer remains to be the embedded word w_p. However, the lower layer's hidden state h_p^{l-1} becomes the input to a cell from the second layer on. Importantly, the hidden state and cell state are updated at each layer with every new bottom-up input to the model.

A competing model that has been shown to slightly outperform the continuously updating ("vanilla") LSTM in character-level language modelling is the hierarchical multiscale LSTM (HM-LSTM (*42*)). This sparsely-updating HM-LSTM, employs a revised updating rule where information from the lower layer is only fed forward at the end of an event. The end of an event is marked by a boundary detector when situational features change. Here, we used the simplified version of the HM-LSTM (*87*) with no top-down connections. See Supplementary Text S2 for a detailed description of the model architecture including all relevant formulas.

696 **Prediction of the next word.** LSTM and HM-LSTM cells form the representations of information
697 relevant to speech prediction, whereas the actual prediction of the next word takes place in the
698 output module. Here, hidden states at word position *p* are combined across the different layers of
699 the language model. The combined hidden state h_p^r is mapped to a fully connected dense layer of
700 as many neurons as there are words in the vocabulary and squashed to values in the interval [0,1],
701 which sum to 1 (i.e., softmax function). Each neuron in resulting vector d_p indexes one particular
702 word in vocabulary *v* and denotes its probability of being the next word. Finally, the word referring
703 to the highest probability in the distribution is chosen as the predicted next word $s_p$in a story. See
704 Supplementary Text S3 for a detailed description of word prediction including all relevant formulas.

705 **Training and evaluation of language models.** The objective of our language models was to
706 minimize the difference between the "predicted" probability distribution d_p (i.e., a vector of
707 probabilities ranging from 0 to 1) and the "actual" probability distribution corresponding to the next
708 word in a text (i.e., a vector of zeros with a one-hot encoded target word). To this end, we trained
709 models on mini-batches of 16 independent text sequences à 500 words and monitored model
710 performance by means of categorical cross-entropy between the "predicted" and "actual"
711 probability distribution of each word in a sequence. Based on model performance, trainable
712 parameters were updated after each mini batch using the Adam algorithm for stochastic gradient
713 optimization (*88*).
714 Our text corpus comprised more than 130 million words including 4,400 political speeches (*89*)
715 as well as 1,200 fictional and popular scientific books. All texts had at least 500 words; metadata,
716 page numbers, references and punctuations (except for hyphenated compound words) were
717 removed from documents. A held-out set of 10 randomly selected documents was used for
718 validation after each epoch of training (i.e., going through the complete training set once) and
719 allowed us to detect overfitting on the training set. Training automatically stopped after model
720 performance did not increase over two epochs for the validation data set.
721 Using a context window of 500 words, we aimed at roughly modelling timescales of the length
722 of common linguistic units in written language (i.e., words, phrases, sentences, and paragraphs).
723 Therefore, we only used a small range of values from three to seven to find the number of layers—
724 intended to represent distinct timescales—best suited to make good predictions. Additionally, we
725 tuned the number of units in LSTM and HM-LSTM cells of language models, using values from 50 to
726 500 in steps of 50. Hyperparameters were evaluated on a single epoch using grid search and the best
727 combination of hyperparameters was chosen based on performance on the validation set. Our final
728 language models had five LSTM or HM-LSTM layers à 300 units and an output module. The LSTM
729 model had 31,428,745 and the HM-LSTM model had 31,431,570 trainable parameters. Models were

trained and evaluated with custom scripts in TensorFlow 2.1. See Supplementary Text S4 for a detailed description of architectural choices.

Deriving the predictiveness of timescales by "masking" language models. We used each trained language model to determine the predictiveness of semantic context in the story presented to participants in the behavioural and fMRI experiment. First, predictiveness was read out from "full" models: We iteratively selected each word in the story as a target word and fed all 500 context words preceding the target word to our language models. Note that the context for target words in the very beginning of the story comprised less than 500 words. The "predicted" probability of each word in the vocabulary was extracted from distribution $\mathrm{d_p}$ in the output module.

Second, predictiveness was read out from "masked" models, where we allowed information to freely flow through networks, yet only considered semantic context represented at single layers to generate the "predicted" probability distribution. These timescale-resolved probabilities were created by setting weight matrix W_r^l of pre-trained models to zero for all layers of no interest, so that the hidden state of only one layer is passed to the softmax function and all other layers have no bearing on the final prediction. We iteratively set all but one layer to zero with each layer being the only one influencing predictions once, resulting in five masked outputs for each language model.

We derived three measures of predictiveness from probability distributions. Our primary measure was the degree of surprisal associated with the occurrence of a word given its context. Word surprisal is the negative logarithm of the probability assigned to the actual next word in a story.

Secondary measures of predictiveness were used to explore the specificity of the processing hierarchy to only some aspects of the prediction processes. Entropy reflects the amount of uncertainty across the whole probability distribution of candidate words, which is the negative sum of probabilities multiplied by their natural logarithm. When high probabilities are assigned to only one or few words in the vocabulary, entropy is low. On the other hand, entropy is high when semantic context is not informative enough to narrow predictions down to a limited set of words, resulting in similar probabilities for all candidate words. As all information necessary to determine the entropy of a word is already available to participants before word presentation, entropy of word w_p was ascribed to the previous word w_{p-1}. Whereas entropy quantifies the overall difficulty of making any definite prediction, surprisal quantifies the availability of information on the actual next word.

Another secondary measure of predictiveness was the relatedness of the predicted next word to the actual next word. This word dissimilarity is expressed as 1 minus the correlation of respective word embeddings. A high positive Product-moment correlation indicates that the prediction is linguistically close to the target word, even though the model prediction might have been incorrect.

764 All three measures were calculated for each word in the story, separately for full models and five
765 masked models. This yielded an 18-dimensional feature space of predictiveness for the LSTM as well
766 as the HM-LSTM model, which was linked to BOLD activity and reading times in our analysis.
767 Additionally, we created a metric to dissociate neural effects of predictiveness from more low-
768 level effects of semantic incongruence between target words and their preceding context. To this
769 end, we correlated the embedding of each function word in the story with the average embedding
770 of a context window, and subtracted resulting Product-moment correlation coefficients from 1 (*90*).
771 This measure of contextual incongruence was calculated at five timescales corresponding to a
772 logarithmic increase in context length (i.e., 2, 4, 8, 16, and 32 words).
773 To determine temporal integration windows of layers, we scrambled input to language models
774 at nine levels of granularity corresponding to a binary logarithmic increase in the length of intact
775 context (i.e., context windows of 1–256 words). For each layer, we fit linear functions to word
776 surprisal across context windows and extracted slope parameters indicating how much a layer
777 benefits from longer context being available when predicting the next word. On the second level,
778 we fit linear functions to these layer-specific integration windows to determine the context benefit
779 of higher layers over shorter layers. Resulting model-specific slopes were compared to a null
780 distribution of slopes computed by shuffling integration windows across layers (n = 10,000).
781 Additionally, slopes were compared between language models by means of a Monte Carlo
782 approximated permutation test (n = 10,000) on the difference of means.

783 **Convolving features with the hemodynamic response function.** We used three classes of
784 features to model brain responses: 18 features of predictiveness (i.e., surprisal, entropy and
785 dissimilarity for the full model as well as for the five masked models), 3 linguistic features, and 9
786 acoustic features. While we were primarily interested in modelling effects of predictiveness,
787 linguistic and acoustic features were used as nuisance regressors potentially covarying with
788 predictiveness. Linguistic features included information on when words were presented (coded as
789 1), whether they were content or function words (coded as 1 and -1), and which frequency they had.
790 In Ref. (*48*), we decomposed the speech-in-noise stimuli into a 288-dimensional acoustic space of
791 spectral, temporal and spectro-temporal modulations, which was derived from a filter bank
792 modelling auditory processing (*91*). Here, we reduced the number of acoustic features to the first 9
793 principal components, which explained more than 80 % of variance in the original acoustic space.
794 All features were z-scored per run.
795 A set of 500 scrambled features of predictiveness was generated, which was used to estimate null
796 distributions of predictive processing. We applied the fast Fourier transform to single features,
797 randomly shifted the phase of frequency components, and inverted the transform to project the
798 data back into the time domain. This preserved power spectra of features but disrupted the temporal

799 alignment of frequencies. See Supplementary Text S5 for a detailed description of convolving
800 features with the hemodynamic response function (HRF).

801 **Data analysis**

802 See Supplementary Text S6 for a detailed description of structural and functional MRI data
803 preprocessing.

804 **Selection of regions of interest.** We hypothesized that the speech prediction hierarchy is
805 represented as a gradient along a temporo-parietal pathway. This rather coarse region of interest
806 was further refined to only include regions implicated in speech processing. To this end, we used
807 intersubject correlation (*50*) as a measure of neural activity consistently evoked across participants
808 listening to speech in noise. As we were primarily interested in shared responses to the speech
809 stream, this approach allowed us to leverage the inconsistency of the noise stream across
810 participants. The presentation of sound textures in different order likely evoked more
811 heterogeneous neural responses, leading to a diminished shared representation of the noise stream.
812 Therefore, we inferred that the shared neural responses we observed were largely driven by the
813 speech stream, which was the same for all participants.

814 At the first level, hyperaligned functional time series of each participant (see Supplementary Text
815 S6) were concatenated across experimental runs and correlated with every other participant on a
816 vertex-by-vertex basis, resulting in pairwise maps of intersubject Product-moment correlations. A
817 group map was created by calculating the median correlation coefficient across pairs of participants
818 for each vertex. At the second level, median correlation coefficients were chosen as a non-parametric
819 test statistic to account for inflated false positive rates and a bootstrap hypothesis test with 10,000
820 iterations was applied. To create the null distribution, we iteratively resampled participants with
821 replacement and derived median group maps from their pairwise correlation maps. When the same
822 participant was sampled more than once in a bootstrap iteration, the pairwise correlation map of
823 that participant with herself was not included in the computation of the group map. The actual
824 median intersubject correlation was ranked against the normalized null distribution to obtain a *p*-
825 value for each vertex. Intersubject correlations were computed with the Python package BrainIAK
826 (*92*) following the recommendations of Nastase and colleagues (*50*).

827 Finally, we used a multi-modal parcellation (*52*) to select those lateral temporal and parietal
828 parcels of which at least 80 % of the vertices had a significant intersubject correlation in one
829 hemisphere ($p < 0.01$, adjusted for false discovery rate; FDR (*93*)). The following parcels were included
830 in the region of interest (ROI): early auditory cortex (EAC), auditory association cortex (AAC), lateral
831 temporal cortex (LTC), temporo-parietal-occipital junction (TPOJ), inferior parietal cortex (IPC), and
832 superior parietal cortex (SPC). As the temporal MT+ complex is thought to be mainly involved in

visual processing, this region was not considered an appropriate candidate parcel. All further analyses including MRI data were limited to the temporo-parietal ROI, which was organized along the anterior-posterior (left: 124 mm, right: 167 mm) and inferior-superior axis (left: 234 mm, right: 212 mm).

Functional data analysis. The starting point of our analyses was the question whether the timescales of speech prediction organize along a temporo-parietal processing hierarchy. In a forward model, we encoded the predictiveness of timescales into univariate neural activity and fit a gradient along the peak locations sensitive to surprisal at specific timescales. Next, we compared the explanatory power of both language models in a supplementary backward model, which decoded surprisal at different timescales from multivariate patterns of neural activity in temporo-parietal parcels. Finally, we modelled functional connectivity between peak locations to test whether the timescales of surprisal gate the information flow along the gradient.

Encoding model. The encoding approach (similar to e.g., Ref. (*94*)) allowed us to quantify for each temporo-parietal vertex, which features of predictiveness it preferentially represents. Two separate encoding models were estimated for each vertex in the ROI of single participants, one for each language model. Besides the features of predictiveness specific to language models, both models included the same linguistic and acoustic features as nuisance regressors. We modelled neural activity as a function of the HRF-convolved features characterizing speech and noise stimuli by:

$$a = Sw + \epsilon,$$

where $a^{samples \times 1}$ is the activity vector (or BOLD time course) corresponding to a vertex, $S^{samples \times features}$ is the stimulus matrix of features, $w^{features \times 1}$ is a vector of estimated model weights, and $\epsilon^{samples \times 1}$ is a vector of random noise.

All models were estimated using ridge regression with four-fold cross validation. We paired odd-numbered functional runs with their subsequent even-numbered run, resulting in four data splits per participant. Each of the four data splits was selected as a testing set once; all other data splits were used as a training set. Within each fold, generalized cross-validation (*95*) was carried out on the training set to find an optimal estimate of regularization parameter λ from the data, searching 100 values evenly spaced on a logarithmic scale from 10^{-5} to 10^{8}. Weights of predictiveness were extracted from the model fit with the optimal regularization parameter and averaged across cross-validation folds to obtain stable weights.

To evaluate the performance of encoding models and their ability to generalize to new data, we applied the weights estimated on the training set to the features of the held-out testing set in each cross-validation fold. The predicted BOLD time series was correlated with the actual BOLD time

series. The resulting Product-moment correlation coefficient is the encoding accuracy, which was averaged across cross-validation folds and Fisher z-transformed.

Additionally, we created null distributions of weights and encoding accuracies by estimating forward models on scrambled features of predictiveness (similar to e.g., Ref. (*96*)). We set up 500 separate models, which included scrambled features of predictiveness but intact linguistic and acoustic features. Models were estimated largely following the cross-validation scheme outlined for observed data. However, we re-used optimal regularization parameters from non-scrambled models of corresponding folds. All ridge regression models were implemented using the RidgeCV function in the Python package scikit-learn (*97*).

Peak selection. For both language models, we derived five temporo-parietal maps in the left and right hemisphere of single participants: one weight map for each timescale of word surprisal. Maps represented the sensitivity of brain regions to timescale surprisal; positive weights indicate increasing BOLD activity to more surprising words.

To illustrate the location and extent of brain regions modulated by timescale surprisal, we performed an analysis similar to cluster-based permutation tests in Fieldtrip (*98*). For each timescale, vertex-wise weights observed across participants were tested against zero by means of a one-sample *t*-test. We combined a vertex into a cluster with its adjacent vertices if it was significant at an alpha level of 0.05 and had at least two significant neighbours. We clustered vertices with negative *t*-values separately from vertices with positive *t*-values. The summed *t*-value of an observed cluster served as the cluster-level statistic and was compared with a Monte Carlo approximated null distribution of summed *t*-values. This null distribution was created by performing clustering on scrambled partitions of timescale-specific weight maps and selecting the largest summed *t*-value for each partition. An observed cluster was considered significant if its summed *t*-value was exceeded by no more than 2.5 % of summed *t*-values from scrambled partitions.

Beyond this rather coarse mapping of temporo-parietal brain regions onto the timescales of surprisal, our main analysis focused on how timescale-specific peak locations distribute along the inferior-superior axis only. Of note, we hypothesized that a hierarchy of speech prediction evolves from temporal to parietal areas, which corresponds to the inferior-superior axis of our ROI. A window with a height of 2 mm was shifted along the inferior-superior axis of the temporo-parietal ROI in steps of 1 mm. All weights of a timescale falling into the window were averaged, thereby collapsing across the anterior-posterior axis. The resulting one-dimensional weight profile of timescale surprisal spanned inferior to superior locations and was smoothed using robust linear regression over a window of 70 mm. For each unilateral weight profile of single participants, local maxima (i.e., a sample larger than its two neighbouring samples) were determined.

We applied two different approaches to select one peak location for each timescale from these local maxima. In the unconstrained approach of peak selection, the local maximum with the highest

902 positive value was defined as a peak. As this approach makes it hard to find a consistent order of
903 timescales when surprisal is not just processed along the dorsal but also the ventral processing
904 stream, we also applied a constrained peak selection approach explicitly targeting the dorsal stream.
905 Here, the peak of surprisal at the first timescale had to be in the inferior half of the axis (i.e., temporal
906 regions) and peaks of surprisal at longer timescales had to be superior to the peak of the first
907 timescale. Whenever no timescale peak could be defined, the largest positive value was selected.
908 Both peak selection approaches yielded five timescale-specific coordinates on the inferior-superior
909 axis for each participant, hemisphere and language model.

910 ***Gradient fitting.*** We fit linear functions to coordinates of single participants across the timescales
911 of surprisal. Models included an intercept term and the slope parameter was extracted from each fit.
912 A positive slope indicates a gradient of timescale surprisal, where surprisal at shorter timescales is
913 represented in more (inferior) temporal regions than longer timescale, which are represented in
914 more (superior) parietal regions. We tested grand-average slope parameters against a null
915 distribution of slopes with 10,000 partitions, which was created by randomly shuffling the
916 coordinates of single participants across the timescales of surprisal and recalculating their slopes. As
917 the first timescale was pre-set to have the most inferior coordinate in the constrained peak selection
918 approach, this specific coordinate was not shuffled when calculating the null distribution for this
919 approach. To compare slope parameters between language models, we performed a Monte Carlo
920 approximated permutation test with 10,000 iterations, using the difference of means as a test
921 statistic.

922 As secondary analyses, gradients of predictive processing were also calculated for the timescales
923 of word entropy and dissimilarity. In a control analysis, a gradient was fit to timescale peaks following
924 the same procedure described above but replacing features of predictiveness by contextual
925 incongruence when estimating forward models.

926 Additionally, we applied unconstrained peak selection to weight maps, which were rotated
927 around the full circle in steps of 5° before collapsing across the first dimension. After deriving test
928 statistics for slope parameters at each rotation as described for non-rotated slope parameters,
929 cluster-based permutation tests were calculated following a similar procedure as described for
930 timescale-specific maps of regression weights before but this time performing clustering in the
931 rotation domain.

932 To round off the encoding analysis, we compared temporo-parietal encoding accuracies between
933 both language models. As we were interested in effects specific to the predictiveness of speech,
934 encoding accuracies were z-scored to the null distribution of accuracies from scrambled features of
935 predictiveness. A cluster-based permutation paired-sample *t*-test was calculated (n = 1,000, vertex-
936 specific alpha level: 0.05, cluster-specific alpha level: 0.05). In comparison to the cluster test described

937 above for the weight maps, we here created a null distribution of summed *t*-values by contrasting
938 accuracies of language models whose labels had been randomly shuffled in single participants.

939 ***Functional connectivity.*** To model the information flow between brain regions sensitive to
940 surprisal at the different timescales, we identified five unique seeds for both language models in
941 each temporo-parietal hemisphere. On the inferior-superior axis, we re-used the grand-median
942 coordinate of each timescale as localized in the constrained peak selection approach. The
943 corresponding coordinate on the anterior-posterior axis was localized by shifting a moving average
944 with a window centred on the inferior-superior coordinate along the anterior-posterior axis (width:
945 2 mm, height: 5 mm), and determining peak locations on smoothed weight profiles of single
946 participants. Following, we placed a sphere with a radius of 5 mm on peak coordinates from both
947 axes and averaged BOLD time courses of vertices falling within this sphere, yielding the timescale-
948 specific neural activity of seeds.

949 We expected increased information flow between seeds of adjacent timescales when one
950 timescale becomes uninformative for the prediction of upcoming speech. This modulatory influence
951 of surprisal on connectivity was modelled along the lines of a psychophysiological interaction (PPI
952 (*53*)). In a standard PPI analysis, the neural time series of one brain region is regressed onto the
953 pointwise product of an experimental stimulus and the neural time series of another brain region.
954 Here, we extended this approach by creating timescale-specific interactions: BOLD time series of
955 seeds were multiplied by their corresponding HRF-convolved surprisal time series but not any of the
956 surprisal time series at another timescale.

957 Functional connectivity was calculated for both language models in each participant and
958 hemisphere. We set up five regression models, with every seed being selected as a target once. The
959 physiological (BOLD) time series of the target seed was mapped onto the physiological,
960 psychological (timescale-specific surprisal), and psychophysiological time series from all other
961 (predictor) seeds. Models were estimated within the same cross-validation scheme outlined for the
962 encoding model. We extracted all four weights from psychophysiological interaction terms of each
963 target seed and arranged weights in a 5-by-5 matrix, with target seeds on the main diagonal and
964 predictor seeds off the diagonal. This matrix of observed psychophysiological interactions was
965 compared to a matrix with hypothesized interaction weights: The diagonals below and above the
966 main diagonal were set to 1 (indicating increased coupling when surprisal at a neighbouring
967 timescale is high), all other items were set to -1. We calculated the Euclidean distance of single-
968 participant matrices to this hypothesized matrix. The mean of observed Euclidean distances was
969 compared to a null distribution of 10,000 mean Euclidean distances calculated on BOLD time series
970 of target seeds randomly shifted in time by the number of samples in 1 to 7 functional runs. Euclidean
971 distances were compared between language models in each hemisphere by means of a Monte Carlo
972 approximated permutation test (n = 10,000) on the difference of means.

973 **Behavioural data analysis.** Reading times were used to test the behavioural relevance of the
974 predictiveness determined by our language models. Trials with reading times shorter than 0.001
975 seconds or longer than 6 seconds were considered invalid and excluded. Further, we inverted
976 reading times into speed and excluded trials exceeding 3 standard deviations within a run and
977 participant from all further analyses. On average, 1.31 % of trials (Ra = 0.32–6.15 %, SD = 1.12 %) were
978 removed. Participants performed the task at a mean response speed of 4.96 s^{-1} (Ra = 1.71–10.13 s^{-1},
979 SD = 2.08 s^{-1}), with speed increasing over the four experimental runs ($p < 0.001$, $d = 1.06$; compared
980 to a null distribution of slopes drawn from linear functions fit to shuffled mean speed of runs). Finally,
981 reading speed was z-scored within runs.

982 For each participant, we predicted reading speed in a forward model, adopting the same cross-
983 validated ridge regression scheme used for the analysis of fMRI data. Our feature space included the
984 predictiveness of words as well as a set of nuisance regressors, namely word frequency, word length
985 (number of letters), content versus function words and trial number. To account for spill-over effects
986 from previous words on reading times (*99*), we added time-lagged versions of features to the model.
987 As previous studies controlled for the spill-over effect on up to three words following (*100*), we
988 shifted features by -2 to 5 word positions to illustrate the effect also returning to zero. There were no
989 lagged versions of the predictor coding for trial number added to the model.

990 To investigate whether predictiveness had an effect on reading speed beyond the effect of
991 nuisance regressors, we compared the predictive accuracy of forward models in single participants
992 to a null distribution of accuracies from models with scrambled features of predictiveness. The
993 performance of language models was compared by z-scoring observed encoding accuracies to the
994 null distribution and running a Monte Carlo approximated permutation test (n = 10,000) on the
995 difference of means. This analysis was also carried out for the timescales of contextual dissimilarity.

References

1. A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*. **36**, 181–204 (2013).

2. J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, C. Padoa-Schioppa, T. Pasternak, H. Seo, D. Lee, X.-J. Wang, A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci*. **17**, 1661–1663 (2014).

3. W. Zhang, M. M. Yartsev, Correlated Neural Activity across the Brains of Socially Interacting Bats. *Cell*. **178**, 413–428 (2019).

4. G. La Camera, A. Rauch, D. Thurbon, H.-R. Lüscher, W. Senn, S. Fusi, Multiple Time Scales of Temporal Response in Pyramidal and Fast Spiking Cortical Neurons. *Journal of Neurophysiology*. **96**, 3448–3464 (2006).

5. J. B. Burt, M. Demirtaş, W. J. Eckner, N. M. Navejar, J. L. Ji, W. J. Martin, A. Bernacchia, A. Anticevic, J. D. Murray, Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat Neurosci*. **21**, 1251–1259 (2018).

6. P. Lakatos, A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, C. E. Schroeder, An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *Journal of Neurophysiology*. **94**, 1904–1911 (2005).

7. M. G. Mattar, D. A. Kahn, S. L. Thompson-Schill, G. K. Aguirre, Varying Timescales of Stimulus Integration Unite Neural Adaptation and Prototype Formation. *Current Biology*. **26**, 1669–1676 (2016).

8. V. A. F. Lamme, P. R. Roelfsema, The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*. **23**, 571–579 (2000).

9. U. Hasson, J. Chen, C. J. Honey, Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Sciences*. **19**, 304–313 (2015).

10. G. T. Buračas, A. M. Zador, M. R. DeWeese, T. D. Albright, Efficient Discrimination of Temporal Patterns by Motion-Sensitive Neurons in Primate Visual Cortex. *Neuron*. **20**, 959–969 (1998).

11. C. A. Runyan, E. Piasini, S. Panzeri, C. D. Harvey, Distinct timescales of population coding across cortex. *Nature*. **548**, 92–96 (2017).

12. R. Chaudhuri, K. Knoblauch, M.-A. Gariel, H. Kennedy, X.-J. Wang, A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron*. **88**, 419–431 (2015).

13. M. Demirtaş, J. B. Burt, M. Helmer, J. L. Ji, B. D. Adkinson, M. F. Glasser, D. C. Van Essen, S. N. Sotiropoulos, A. Anticevic, J. D. Murray, Hierarchical Heterogeneity across Human Cortex Shapes Large-Scale Neural Dynamics. *Neuron*. **101**, 1181-1194.e13 (2019).

14. J. M. Huntenburg, P.-L. Bazin, D. S. Margulies, Large-Scale Gradients in Human Cortical Organization. *Trends in Cognitive Sciences*. **22**, 21–31 (2018).

15. K. D. Himberger, H.-Y. Chien, C. J. Honey, Principles of Temporal Processing Across the Cortical Hierarchy. *Neuroscience*. **389**, 161–174 (2018).

1034 16. K. Friston, A theory of cortical responses. *Phil. Trans. R. Soc. B.* **360**, 815–836 (2005).

1035 17. G. B. Keller, T. D. Mrsic-Flogel, Predictive Processing: A Canonical Cortical Computation.
1036 *Neuron*. **100**, 424–435 (2018).

1037 18. S. J. Kiebel, J. Daunizeau, K. J. Friston, A Hierarchy of Time-Scales and the Brain. *PLoS*
1038 *Computational Biology*. **4**, e1000209 (2008).

1039 19. A. M. Bastos, J. Vezoli, C. A. Bosman, J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P.
1040 De Weerd, H. Kennedy, P. Fries, Visual Areas Exert Feedforward and Feedback Influences
1041 through Distinct Frequency Channels. *Neuron*. **85**, 390–401 (2015).

1042 20. L. Cocchi, M. V. Sale, L. L Gollo, P. T. Bell, V. T. Nguyen, A. Zalesky, M. Breakspear, J. B.
1043 Mattingley, A hierarchy of timescales explains distinct effects of local inhibition of primary
1044 visual cortex and frontal eye fields. *eLife*. **5**, e15252 (2016).

1045 21. C. Wacongne, E. Labyt, V. van Wassenhove, T. Bekinschtein, L. Naccache, S. Dehaene,
1046 Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of*
1047 *the National Academy of Sciences*. **108**, 20754–20759 (2011).

1048 22. C. M. Schwiedrzik, W. A. Freiwald, High-Level Prediction Signals in a Low-Level Area of the
1049 Macaque Face-Processing Hierarchy. *Neuron*. **96**, 89-97.e4 (2017).

1050 23. P. W. Donhauser, S. Baillet, Two Distinct Neural Timescales for Predictive Speech Processing.
1051 *Neuron*. **105**, 385-393.e9 (2020).

1052 24. Z. C. Chao, K. Takaura, L. Wang, N. Fujii, S. Dehaene, Large-Scale Cortical Networks for
1053 Hierarchical Prediction and Prediction Error in the Primate Brain. *Neuron*. **100**, 1252-1266.e3
1054 (2018).

1055 25. G. J. Stephens, C. J. Honey, U. Hasson, A place for time: the spatiotemporal structure of
1056 neural dynamics during natural audition. *Journal of Neurophysiology*. **110**, 2019–2026 (2013).

1057 26. W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, F. E. Theunissen, The Hierarchical Cortical
1058 Organization of Human Speech Processing. *The Journal of Neuroscience*. **37**, 6539–6557
1059 (2017).

1060 27. Y. Lerner, C. J. Honey, L. J. Silbert, U. Hasson, Topographic Mapping of a Hierarchy of
1061 Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*. **31**, 2906–
1062 2915 (2011).

1063 28. C. H. C. Chang, C. Lazaridi, Y. Yeshurun, K. A. Norman, U. Hasson, Relating the Past with the
1064 Present: Information Integration and Segregation during Ongoing Narrative Processing.
1065 *Journal of Cognitive Neuroscience*. **33**, 1106–1128 (2021).

1066 29. I. Bornkessel-Schlesewsky, M. Schlesewsky, S. L. Small, J. P. Rauschecker, Neurobiological
1067 roots of language in primate audition: common computational properties. *Trends in*
1068 *Cognitive Sciences*. **19**, 142–150 (2015).

1069 30. G. R. Kuperberg, T. F. Jaeger, What do we mean by prediction in language comprehension?
1070 *Language, Cognition and Neuroscience*. **31**, 32–59 (2016).

1071 31. L. H. Arnal, V. Wyart, A.-L. Giraud, Transitions in neural oscillations reflect prediction errors
1072 generated in audiovisual speech. *Nat Neurosci*. **14**, 797–801 (2011).

1073 32. H. Blank, M. H. Davis, Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI
1074 Patterns during Speech Perception. *PLoS Biol.* **14**, e1002577 (2016).

1075 33. K. D. Kandylaki, A. Nagels, S. Tune, T. Kircher, R. Wiese, M. Schlesewsky, I. Bornkessel-
1076 Schlesewsky, Predicting "When" in Discourse Engages the Human Dorsal Auditory Stream:
1077 An fMRI Study Using Naturalistic Stories. *Journal of Neuroscience*. **36**, 12180–12191 (2016).

1078 34. K. Friston, S. Kiebel, Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B.*
1079 **364**, 1211–1221 (2009).

1080 35. J. Hale, in *Proceedings of the 2nd North American Chapter of the Association for Computational*
1081 *Linguistics* (Pittsburgh, Pennsylvania, 2001;
1082 http://portal.acm.org/citation.cfm?doid=1073336.1073357), pp. 159–166.

1083 36. R. Levy, Expectation-based syntactic comprehension. *Cognition*. **106**, 1126–1177 (2008).

1084 37. R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, A. van den Bosch, Prediction During
1085 Natural Language Comprehension. *Cerebral Cortex*. **26**, 2506–2516 (2016).

1086 38. I. F. Monsalve, S. L. Frank, G. Vigliocco, Lexical surprisal as a general predictor of reading
1087 time. *Proceedings of the 13th Conference of the European Chapter of the Association for*
1088 *Computational Linguistics*, 398–408 (2012).

1089 39. H.-Y. S. Chien, C. J. Honey, Constructing and Forgetting Temporal Context in the Human
1090 Cerebral Cortex. *Neuron*. **106**, 675-686.e11 (2020).

1091 40. A. Zadbood, J. Chen, Y. C. Leong, K. A. Norman, U. Hasson, How We Transmit Memories to
1092 Other Brains: Constructing Shared Neural Representations Via Communication. *Cerebral*
1093 *Cortex*. **27**, 4988–5000 (2017).

1094 41. C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, K. A. Norman, Discovering Event
1095 Structure in Continuous Narrative Perception and Memory. *Neuron*. **95**, 709-721.e5 (2017).

1096 42. J. Chung, S. Ahn, Y. Bengio, in *Proceedings of the 5th International Conference on Learning*
1097 *Representations* (Toulon, France, 2016; http://arxiv.org/abs/1609.01704).

1098 43. J. D. Cohen, N. Daw, B. Engelhardt, U. Hasson, K. Li, Y. Niv, K. A. Norman, J. Pillow, P. J.
1099 Ramadge, N. B. Turk-Browne, T. L. Willke, Computational approaches to fMRI analysis. *Nat*
1100 *Neurosci*. **20**, 304–313 (2017).

1101 44. R. M. Cichy, D. Kaiser, Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*.
1102 **23**, 305–317 (2019).

1103 45. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Computation*. **9**, 1735–1780
1104 (1997).

1105 46. S. Jain, A. Huth, Incorporating Context into Language Encoding Models for fMRI. *bioRxiv*
1106 (2018), doi:10.1101/327601.

1107 47. C. Aurnhammer, S. L. Frank, Evaluating information-theoretic measures of word prediction in
1108 naturalistic sentence reading. *Neuropsychologia*. **134**, 107198 (2019).

1109 48. J. Erb, L.-M. Schmitt, J. Obleser, Temporal selectivity declines in the aging human auditory
1110 cortex. *eLife*. **9**, e55300 (2020).

1111 49. K. Rayner, Eye Movements in Reading and Information Processing: 20 Years of Research.
1112 *Psychological Bulletin*. **124**, 372–422 (1998).

1113 50. S. A. Nastase, V. Gazzola, U. Hasson, C. Keysers, Measuring shared responses across subjects
1114 using intersubject correlation. *Social Cognitive and Affective Neuroscience*, nsz037 (2019).

1115 51. J. R. Binder, R. H. Desai, The neurobiology of semantic memory. *Trends in Cognitive Sciences*.
1116 **15**, 527–536 (2011).

1117 52. M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J.
1118 Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, D. C. Van Essen, A multi-modal
1119 parcellation of human cerebral cortex. *Nature*. **536**, 171–178 (2016).

1120 53. K. J. Friston, C. Buechel, G. R. Fink, J. Morris, E. Rolls, R. J. Dolan, Psychophysiological and
1121 Modulatory Interactions in Neuroimaging. *NeuroImage*. **6**, 218–229 (1997).

1122 54. U. Hasson, E. Yang, I. Vallines, D. J. Heeger, N. Rubin, A Hierarchy of Temporal Receptive
1123 Windows in Human Cortex. *Journal of Neuroscience*. **28**, 2539–2550 (2008).

1124 55. J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, Event perception: A mind-
1125 brain perspective. *Psychological Bulletin*. **133**, 273–293 (2007).

1126 56. G. A. Radvansky, Across the Event Horizon. *Curr Dir Psychol Sci*. **21**, 269–272 (2012).

1127 57. J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L.
1128 Buckner, M. E. Raichle, Human brain activity time-locked to perceptual event boundaries.
1129 *Nat Neurosci*. **4**, 651–655 (2001).

1130 58. C. Whitney, W. Huber, J. Klann, S. Weis, S. Krach, T. Kircher, Neural correlates of narrative
1131 shifts during auditory story comprehension. *NeuroImage*. **47**, 360–366 (2009).

1132 59. L. L. Richmond, J. M. Zacks, Constructing Experience: Event Models from Perception to
1133 Action. *Trends in Cognitive Sciences*. **21**, 962–980 (2017).

1134 60. F. Lieder, K. E. Stephan, J. Daunizeau, M. I. Garrido, K. J. Friston, A Neurocomputational Model
1135 of the Mismatch Negativity. *PLoS Comput Biol*. **9**, e1003288 (2013).

1136 61. A. Todorovic, F. P. de Lange, Repetition Suppression and Expectation Suppression Are
1137 Dissociable in Time in Early Auditory Evoked Fields. *Journal of Neuroscience*. **32**, 13389–
1138 13395 (2012).

1139 62. P. Hagoort, G. Baggio, R. M. Willems, in *The cognitive neurosciences* (MIT Press, Cambridge,
1140 MA, ed. 4, 2009), pp. 819–836.

1141 63. P. Kok, J. F. M. Jehee, F. P. de Lange, Less Is More: Expectation Sharpens Representations in
1142 the Primary Visual Cortex. *Neuron*. **75**, 265–270 (2012).

1143 64. R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of
1144 some extra-classical receptive-field effects. *Nat Neurosci*. **2**, 79–87 (1999).

1145 65. G. Hickok, D. Poeppel, The cortical organization of speech processing. *Nature Reviews*
1146 *Neuroscience*. **8**, 393–402 (2007).

1147 66. I. DeWitt, J. P. Rauschecker, Phoneme and word recognition in the auditory ventral stream.
1148 *Proceedings of the National Academy of Sciences*. **109**, E505–E514 (2012).

1149 67. I. Bornkessel-Schlesewsky, M. Schlesewsky, Reconciling time, space and function: A new
1150 dorsal–ventral stream model of sentence comprehension. *Brain and Language*. **125**, 60–76
1151 (2013).

1152 68. S. M. Wilson, I. Molnar-Szakacs, M. Iacoboni, Beyond Superior Temporal Cortex: Intersubject
1153 Correlations in Narrative Speech Comprehension. *Cerebral Cortex*. **18**, 230–242 (2008).

1154 69. H. N. Phillips, A. Blenkmann, L. E. Hughes, S. Kochen, T. A. Bekinschtein, Cam-CAN, J. B. Rowe,
1155 Convergent evidence for hierarchical prediction networks from human electrocorticography
1156 and magnetoencephalography. *Cortex*. **82**, 192–205 (2016).

1157 70. B. Lyu, J. Ge, Z. Niu, L. H. Tan, J.-H. Gao, Predictive Brain Mechanisms in Sound-to-Meaning
1158 Mapping during Speech Processing. *J. Neurosci*. **36**, 10813–10822 (2016).

1159 71. F. Meyniel, S. Dehaene, Brain networks for confidence weighting and hierarchical inference
1160 during probabilistic learning. *Proc Natl Acad Sci USA*. **114**, E3859–E3868 (2017).

1161 72. V. K. M. Cheung, L. Meyer, A. D. Friederici, S. Koelsch, The right inferior frontal gyrus
1162 processes nested non-local dependencies in music. *Scientific Reports*. **8**, 3822 (2018).

1163 73. L. H. Tan, A. R. Laird, K. Li, P. T. Fox, Neuroanatomical correlates of phonological processing of
1164 Chinese characters and alphabetic words: A meta-analysis. *Hum. Brain Mapp*. **25**, 83–91
1165 (2005).

1166 74. D. Ayyash, S. Malik-Moraleda, J. Gallée, J. Affourtit, M. Hoffman, Z. Mineroff, O. Jouravlev, E.
1167 Fedorenko, The universal language network: A cross-linguistic investigation spanning 45
1168 languages and 11 language families. *bioRxiv* (2021), doi:10.1101/2021.07.28.454040.

1169 75. R. Bottini, C. F. Doeller, Knowledge Across Reference Frames: Cognitive Maps and Image
1170 Spaces. *Trends in Cognitive Sciences*. **24**, 606–619 (2020).

1171 76. I. K. Brunec, M. Moscovitch, M. D. Barense, Boundaries Shape Cognitive Representations of
1172 Spaces and Events. *Trends in Cognitive Sciences*. **22**, 637–650 (2018).

1173 77. A. S. Alexander, D. A. Nitz, Spatially Periodic Activation Patterns of Retrosplenial Cortex
1174 Encode Route Sub-spaces and Distance Traveled. *Current Biology*. **27**, 1551-1560.e4 (2017).

1175 78. K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, The hippocampus as a predictive map. *Nat*
1176 *Neurosci*. **20**, 1643–1653 (2017).

1177 79. J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, Spurious but systematic
1178 correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*.
1179 **59**, 2142–2154 (2012).

1180 80. A. U. Rysop, L.-M. Schmitt, J. Obleser, G. Hartwigsen, Neural modelling of the semantic
1181 predictability gain under challenging listening conditions. *Human Brain Mapping*. **42**, 110–
1182 127 (2021).

1183 81. J. H. McDermott, E. P. Simoncelli, Sound Texture Perception via Statistics of the Auditory
1184 Periphery: Evidence from Sound Synthesis. *Neuron*. **71**, 926–940 (2011).

1185 82. T. Kisler, U. Reichel, F. Schiel, Multilingual processing of speech via web services. *Computer*
1186 *Speech & Language*. **45**, 326–347 (2017).

83. M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, A. Böhl, The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology*. **58**, 412–424 (2011).

84. W. J. B. van Heuven, P. Mandera, E. Keuleers, M. Brysbaert, Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*. **67**, 1176–1190 (2014).

85. D. H. Brainard, The Psychophysics Toolbox. *Spatial Vis.* **10**, 433–436 (1997).

86. B. Heinzerling, M. Strube, BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. *arXiv:1710.02187 [cs]* (2017) (available at http://arxiv.org/abs/1710.02187).

87. Á. Kádár, M.-A. Côté, G. Chrupała, A. Alishahi, Revisiting the Hierarchical Multiscale LSTM. *arXiv:1807.03595 [cs]* (2018) (available at http://arxiv.org/abs/1807.03595).

88. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017) (available at http://arxiv.org/abs/1412.6980).

89. A. Barbaresi, A corpus of German political speeches from the 21st century. *11th Language Resources and Evaluation Conference*, 792–797 (2018).

90. M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, E. C. Lalor, Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*. **28**, 803-809.e3 (2018).

91. T. Chi, P. Ru, S. A. Shamma, Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America.* **118**, 887–906 (2005).

92. M. Kumar, C. T. Ellis, Q. Lu, H. Zhang, M. Capotă, T. L. Willke, P. J. Ramadge, N. B. Turk-Browne, K. A. Norman, BrainIAK tutorials: User-friendly learning materials for advanced fMRI analysis. *PLoS Comput Biol.* **16**, e1007549 (2020).

93. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B (Methodological)*. **57**, 289–300 (1995).

94. R. Santoro, M. Moerel, F. De Martino, G. Valente, K. Ugurbil, E. Yacoub, E. Formisano, Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc Natl Acad Sci USA.* **114**, 4799–4804 (2017).

95. G. H. Golub, M. Heath, G. Wahba, Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*. **21**, 215–223 (1979).

96. S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, A. K. Churchland, Single-trial neural dynamics are dominated by richly varied movements. *Nat Neurosci.* **22**, 1677–1686 (2019).

97. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. **12**, 2825–2830 (2011).

98. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*. **164**, 177–190 (2007).

1225 99. D. C. Mitchell, in *New methods in reading comprehension research* (Erlbaum, Hillsdale, N.J.,
1226 1984), pp. 69–89.

1227 100. N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic.
1228 *Cognition*. **128**, 302–319 (2013).

1229 101. J. Pennington, R. Socher, C. Manning, in *Proceedings of the 2014 Conference on Empirical*
1230 *Methods in Natural Language Processing* (Doha, Qatar, 2014;
1231 http://aclweb.org/anthology/D14-1162), pp. 1532–1543.

1232 102. J. L. Ba, J. R. Kiros, G. E. Hinton, Layer Normalization. *arXiv:1607.06450 [cs, stat]* (2016)
1233 (available at http://arxiv.org/abs/1607.06450).

1234 103. W. Penny, K. Friston, J. Ashburner, S. Kiebel, T. Nichols, *Statistical Parametric Mapping: The*
1235 *Analysis of Functional Brain Images* (Academic Press, 2006).

1236 104. O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M.
1237 Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, K. J.
1238 Gorgolewski, fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*.
1239 **16**, 111–116 (2019).

1240 105. K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, S. S. Ghosh,
1241 Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in
1242 Python. *Frontiers in Neuroinformatics*. **5** (2011), doi:10.3389/fninf.2011.00013.

1243 106. A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B.
1244 Thirion, G. Varoquaux, Machine learning for neuroimaging with scikit-learn. *Frontiers in*
1245 *Neuroinformatics*. **8** (2014), doi:10.3389/fninf.2014.00014.

1246 107. N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, N4ITK:
1247 Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*. **29**, 1310–1320 (2010).

1248 108. A. M. Dale, B. Fischl, M. I. Sereno, Cortical Surface-Based Analysis. *NeuroImage*. **9**, 179–194
1249 (1999).

1250 109. V. Fonov, A. Evans, R. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-
1251 appropriate brain templates from birth to adulthood. *NeuroImage*. **47**, S102 (2009).

1252 110. B. Avants, C. Epstein, M. Grossman, J. Gee, Symmetric diffeomorphic image registration with
1253 cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain.
1254 *Medical Image Analysis*. **12**, 26–41 (2008).

1255 111. A. Klein, S. S. Ghosh, F. S. Bao, J. Giard, Y. Häme, E. Stavsky, N. Lee, B. Rossa, M. Reuter, E.
1256 Chaibub Neto, A. Keshavan, Mindboggling morphometry of human brains. *PLoS Comput Biol*.
1257 **13**, e1005350 (2017).

1258 112. Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden Markov
1259 random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging*.
1260 **20**, 45–57 (2001).

1261 113. M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved Optimization for the Robust and
1262 Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*. **17**, 825–
1263 841 (2002).

114. R. W. Cox, AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*. **29**, 162–173 (1996).

115. D. N. Greve, B. Fischl, Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*. **48**, 63–72 (2009).

116. R. H. R. Pruim, M. Mennes, D. van Rooij, A. Llera, J. K. Buitelaar, C. F. Beckmann, ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*. **112**, 267–277 (2015).

117. J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*. **84**, 320–341 (2014).

118. M. A. Lindquist, S. Geuter, T. D. Wager, B. S. Caffo, Modular preprocessing pipelines can reintroduce artifacts into fMRI data. *Human Brain Mapping*. **40**, 2358–2376 (2019).

119. J. S. Guntupalli, M. Hanke, Y. O. Halchenko, A. C. Connolly, P. J. Ramadge, J. V. Haxby, A Model of Representational Spaces in Human Cortex. *Cereb. Cortex*. **26**, 2919–2934 (2016).

120. M. Hanke, Y. O. Halchenko, P. B. Sederberg, S. J. Hanson, J. V. Haxby, S. Pollmann, PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinform*. **7**, 37–53 (2009).

121. R. Boldt, S. Malinen, M. Seppä, P. Tikka, P. Savolainen, R. Hari, S. Carlson, Listening to an Audio Drama Activates Two Processing Networks, One for All Sounds, Another Exclusively for Speech. *PLoS ONE*. **8**, e64489 (2013).

122. R. Schmälzle, F. E. K. Häcker, C. J. Honey, U. Hasson, Engaged listeners: shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience*. **10**, 1137–1143 (2015).

123. M. Regev, E. Simony, K. Lee, K. M. Tan, J. Chen, U. Hasson, Propagation of Information Along the Cortical Hierarchy as a Function of Attention While Reading and Listening to Stories. *Cerebral Cortex*. **29**, 4017–4034 (2018).

Acknowledgements

We thank Anne Herrmann, Clara Mergner, Malte Naujokat, Anna Ruhe, and Svenja Meyn for their help with data acquisition; Christine Sickert for her help in preparing the text corpus; Martin Göttlich for setting up the MR sequences; Mattias Heinrich for discussions on natural language processing; and Malte Wöstmann for suggestions to the reading-task design.

Funding:

German Research Foundation grant OB 352/2-1 (JO)

German Research Foundation grant HA 6314/4-1 (GH)

European Research Council consolidator grant ERC-CoG-2014–646696 (JO)

Author contributions:

Conceptualization:	LMS, JE, JO
Data curation:	LMS, JE
Formal analysis:	LMS, JE, ST, JO
Funding acquisition:	GH, JO
Investigation:	LMS, JE
Methodology:	LMS, JE, ST, AR, GH, JO
Project administration:	LMS, JE
Resources:	LMS, JE, JO
Software:	LMS, JE
Supervision:	GH, JO
Validation:	LMS, JO
Visualization:	LMS
Writing – Original draft:	LMS
Writing – Review & editing:	LMS, JE, ST, AR, GH, JO

Competing interests: The authors declare no competing interests.

Data and materials availability: Functional MRI data are publicly available on the Open Science Framework (OSF; https://osf.io/zbuah). All other data needed to evaluate the conclusions in the paper are present in the main text or the supplementary materials.