Finding structure during incremental speech comprehension

Bingjiang Lyu[1*], William D. Marslen-Wilson[2], Yuxing Fang[2], Lorraine K. Tyler[2*]

[1]Changping Laboratory, Beijing, 102206, China

[2]Centre for Speech, Language and the Brain, Department of Psychology,

University of Cambridge, Cambridge, CB2 3EB, United Kingdom

* Correspondence: bingjiang.lyu@gmail.com (BL), lkt10@cam.ac.uk (LKT).

Abstract

A core aspect of human speech comprehension is the ability to incrementally integrate consecutive words into a structured and coherent interpretation, aligning with the speaker's intended meaning. This rapid process is subject to multi-dimensional probabilistic constraints, including both linguistic knowledge and non-linguistic information within specific contexts, and it is their interpretative coherence that drives successful comprehension. To study the neural substrates of this process, we extract word-by-word measures of sentential structure from BERT, a deep language model, which effectively approximates the coherent outcomes of the dynamic interplay among various types of constraints. Using representational similarity analysis, we tested BERT parse depths and relevant corpus-based measures against the spatiotemporally resolved brain activity recorded by electro/magnetoencephalography when participants were listening to the same sentences. Our results provide a detailed picture of the neurobiological processes involved in the incremental construction of structured interpretations. These findings show when and where coherent interpretations emerge through the evaluation and integration of multifaceted constraints in the brain, which engages bilateral brain regions extending beyond the classical fronto-temporal language system. Furthermore, this study provides empirical evidence supporting the use artificial neural networks as computational models for revealing the neural dynamics underpinning complex cognitive processes in the brain.

Keywords: incremental speech comprehension, artificial neural networks, deep language models, representational similarity analysis, computational cognitive neuroscience, EEG/MEG

Introduction

Human speech comprehension involves a complex set of processes that transform an auditory input into the speaker's intended meaning, wherein each word is sequentially recognized and integrated with the preceding words to obtain a coherent interpretation (Marslen-Wilson and Tyler 1980; Lyu et al. 2019; Choi et al. 2021). Crucially, rather than simple linear concatenation, individual words are combined according to the nonlinear and often discontinuous structure embedded in an utterance as it is delivered over time (Everaert et al. 2015). For example, in the sentence *"The boy who chased the cat was…"*, it is the structurally close word *"boy"*, rather than the linearly close word *"cat"*, that is combined with *"was"*. However, the neural dynamics underpinning the incremental construction of a structured interpretation from a spoken sentence is still unclear.

Previous neuroimaging studies on the structure of language primarily focused on syntax (Matchin and Hickok 2020), contrasting grammatical sentences against word lists or sentences with syntactic violations (Nelson et al. 2017; Law and Pylkkanen 2021), manipulating the syntactic complexity in sentences (Pallier et al. 2011), or studying artificial grammatical rules elicited by structured, unintelligible strings (Friederici et al. 2006). Nevertheless, finding the structure in an unfolding sentence also depends on the constraints jointly placed by other linguistic properties and non-linguistic information such as the broad world knowledge (Bever 1970; Tyler and Marslen-Wilson 1977).

Unlike the *two-stage model* (Frazier and Rayner 1982; Frazier 1987) which posits an initial parsing stage relying solely on syntax, the *constraint-based* approach to sentence processing (MacDonald et al. 1994; Trueswell and Tanenhaus 1994) proposes that speech comprehension is concurrently governed by multiple types of probabilistic constraints (e.g., syntax, semantics, world knowledge), generated by individual words as they are sequentially heard. There is no delay in the utilization of these multifaceted constraints once they become available, neither is a fixed priority assigned to one type of constraint over another; rather, it is the *interpretative coherence* of all available constraints that forms the basis for successful language comprehension (Altmann 1998). Although lexical constraints of individual words can be estimated from large corpora data, it has been challenging to model the dynamic interplay between various types of linguistic and nonlinguistic constraints in a specific context, especially at the sentence level and beyond.

Contemporary deep language models (DLMs) have made great strides in a wide array of natural language processing tasks, including text generation, parsing and translation (Vaswani et al. 2017; Devlin et al. 2019; Brown et al. 2020; Ouyang et al. 2022). While current DLMs are still imperfect in terms of human-level language understanding related to reasoning and complex physical or social situations (Bisk et al. 2020), they are arguably valuable models of general linguistic capacities, due to their ability to identify and leverage relevant statistical regularities of linguistic and non-linguistic world knowledge present in massive training data (Linzen and Baroni 2021; Pavlick 2022). Human language comprehension requires a contextualized integration of multifaceted constraints (Marslen-Wilson 1975; Tyler and Marslen-Wilson 1977;

Kuperberg 2007). In this regard, DLMs excel in flexible combination of different types of features (e.g., syntactic structure and semantic meaning) embedded in their rich internal representations (Manning et al. 2020; Bengio et al. 2021; Linzen and Baroni 2021; Pavlick 2022). Their deep contextualized representations capture the distributed regularities that jointly determine the coherent interpretation of a given sentence, providing context-dependent composition and quantitative measures of the underlying sentential structure. These properties relate back to Elman's recurrent neural network (Elman 1990, 1993) which automatically picks up and encodes lexical syntactic/semantic information in the hidden states.

Recent studies have revealed an overall congruence between language representations in DLMs and those observed in the human brain while processing the same spoken or written input (Schrimpf et al. 2021; Caucheteux et al. 2022; Caucheteux and King 2022; Goldstein et al. 2022; Heilbron et al. 2022; Toneva et al. 2022; Caucheteux et al. 2023), suggesting the potential value of DLMs as a computational tool to investigate the neural basis of language comprehension. To move beyond comparing the similarities between entire model hidden states and brain activity, probing techniques that can extract specific contents from DLMs (Hewitt and Liang 2019; Tenney et al. 2019) make it possible to study the neural dynamics relevant to processing such specific information. The important advance here is that we can leverage the deep learning strengths of DLMs to create rigorously quantified models of the broader and multifaceted constraint environment in which a structured interpretation is constructed. Such models can be compared, dynamically, with more restricted and interpretable factors that capture the specific linguistic combinatorial constraints necessary for successful language comprehension.

Here, we take this approach further by designing sentences with contrasting linguistic structures and using a structural probe technique (Hewitt and Manning 2019) to extract word-by-word contextualized representations of sentential structures from a widely-used DLM, namely, BERT (Devlin *et al.* 2019). This provides the neurocomputational specificity required to elucidate the neural dynamics underlying the incremental construction of a structured interpretation from an unfolding spoken sentence. After a detailed evaluation of BERT structural measures according to the hypothesized *constraint-based* approach and human behavioural results, we used spatiotemporal searchlight representational similarity analysis (ssRSA) (Kriegeskorte et al. 2008) to test these quantitative structural measures and relevant lexical properties against source-localized EMEG data recorded while participants were listening to the same sentences. Our findings reveal how the structured interpretation of a spoken sentence is incrementally built under multifaceted probabilistic constraints in the brain.

Results

We constructed 60 sets of sentences with varying sentential structures (see Methods) and presented them to human listeners. We also input them word-by-word to BERT to extract incremental structural representations. These natural spoken sentences were constructed to balance off specifically linguistic constraints on interpretation against varying non-linguistic

constraints as the sentence is incrementally interpreted, providing a realistic simulation of real-life language use. In each stimulus set, there are two target sentences differing only in the transitivity of the first verb (Verb1) encountered, i.e., how likely it is that Verb1 takes a direct object [see (1) and (2) below and **Figure 1**]:

(1) *The dog found in the park was covered in mud.*
(2) *The dog walked in the park was covered in mud.*

In the first sentence, Verb1 (i.e., *"found"*) has high transitivity (HiTrans) and strongly prefers a direct object (e.g., ball), while in the second sentence, Verb1 (i.e., *"walked"*) has relatively low transitivity (LoTrans) and is often used without a following direct object. Critically, (a) the structural interpretation of these sentences is ambiguous at the point Verb1 is encountered and (b) the preferred human resolution of this ambiguity depends on the real-time integration of linguistic and non-linguistic constraints as more of the sentence is heard. In the example above, the sequence *"The dog found...*" could initially have either an Active interpretation – where the dog has found something, or a Passive interpretation – where the dog is found by someone (**Figure 1**). Because *"find"* is primarily a transitive verb, the human listener is likely to be biased towards an initial Active interpretation. Similarly, the sequence *"The dog walked…"*, where *walk* is primarily used as an intransitive verb (without a direct object), could also bias the listener to an Active interpretation, where the dog is doing the walking, rather than the less frequent Passive interpretation where someone is taking the dog for a walk (i.e., walking the dog).

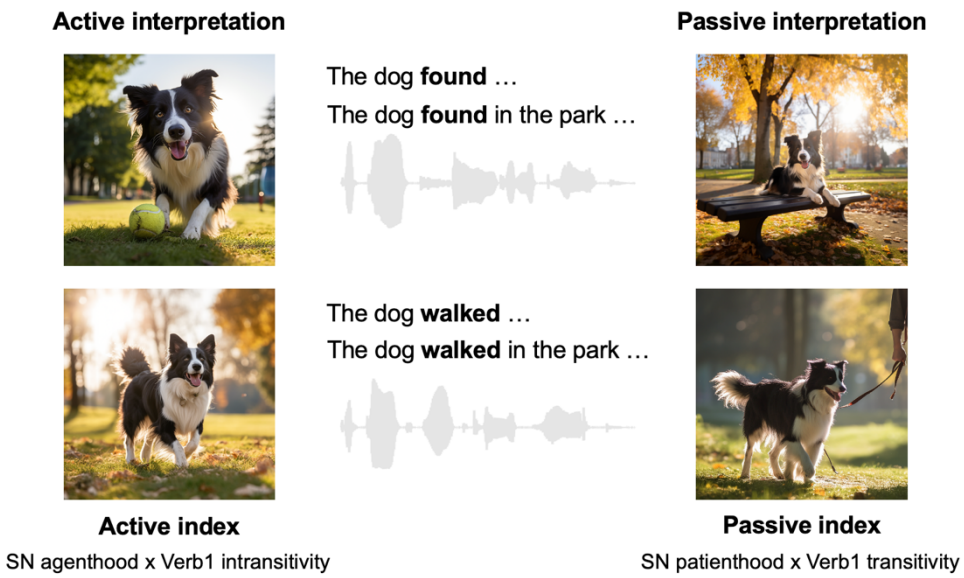

Figure 1. Example spoken sentence stimuli and plausible structured interpretations. The two target sentences in each set differ only in the transitivity of the first verb (Verb1). Each sentence has two possible structured interpretations before the actual main verb is presented: an active interpretation, where the subject noun (SN) performs the action, and a passive interpretation, where the SN is the recipient of the action. The interpretative preference hinges on the likelihood of the SN acting as an agent or a patient (i.e., its thematic role) in conjunction with the transitivity of Verb1. As the sentence progresses to the prepositional phrase, a combination of higher SN agenthood and greater Verb1 intransitivity (i.e., a higher Active index) generally favors an Active interpretation. Conversely, increased SN patienthood coupled with higher Verb1 transitivity (i.e., a higher Passive index) may lead to a Passive interpretation. Note that while the SN is the same for the

two target sentences within the same set, it varies across different sentence sets. All images were generated using Midjourney for illustrative purposes.

This initial structural interpretation up to Verb1 does not, however, just depend on linguistic knowledge such as Verb1 transitivity. It also depends on non-linguistic information, i.e., how likely the subject is (or is not) to adopt the Active (agent) role to perform the specified action (Dowty 1991; Marslen-Wilson et al. 1993), that is, "thematic role" properties of the subject noun. Although it could be reflected by statistical regularities in language, thematic role preference hinges more on world knowledge, plausibility, or real-world statistics. So, regardless of Verb1 transitivity, the Active interpretation should be more strongly favored in *"The **king** found/walked..."* given the higher agenthood of the *"king"* and thus the greater implausibility of a Passive interpretation involving a *"king"* relative to a *"dog"*. Hence, the word-by-word interpretation of the sentential structure – and of the real-world event structure evoked by this interpretation – is determined by the constraints jointly placed by the subject noun and Verb1, which is manifested by the interpretative coherence between non-linguistic world knowledge (i.e., thematic role preference) and linguistic knowledge (i.e., verb transitivity).

As the sentence evolves, and the prepositional phrase *"in the park"* that follows Verb1 is incrementally processed, there is further modulation of the preferred interpretation, again reflecting both Verb1 transitivity and the plausibility of the event being constructed. Specifically, the Passive interpretation will become more preferred in a HiTrans sentence, given the absence of an expected direct object for the highly transitive Verb1, so Verb1 tends to be interpreted as a passive verb [i.e., the head of a reduced relative clause in *"The dog **(that was) found in the park**…"*]. Conversely, in a LoTrans sentence, the Active interpretation of Verb1 is strengthened by the incoming prepositional phrase, which is in accord with the verb's intransitive use and the event conjured up by the sequence of words heard so far (e.g., *"The dog walked in the park…"*). Hence, these two sentence types are likely to differ in the structural interpretation preferred by the end of the prepositional phrase. However, with the appearance of the actual main verb (e.g., *"was covered"* in the example sentences), the Active interpretation of Verb1 as the main verb will be completely rejected, which resolves the potential ambiguity and confirms the Passive interpretation in both HiTrans and LoTrans sentences.

In brief, understanding these complex sentences require listeners to integrate discontinuous words to solve a long-distance dependency between the subject noun and the actual main verb separated by an intervening clause. This engages the neurobiological processes of integration across different lexical constraints and multiple levels of the sentence processing system. For example, the incremental building, maintenance and update of sentential structure over time might primarily involve activity in the fronto-temporal regions (Friederici 2012), while estimating the plausibility of the event interpreted from the sentence with prior knowledge of the world may elicit neural responses in the default mode network (Yeshurun et al. 2021).

Human incremental structural interpretations

As the first step, and to quantify how the stimulus sentences exemplified a constraint-based account of incremental structural interpretation, we conducted two pre-tests where participants listened to sentence fragments, starting from sentence onset and continuing either until the end of Verb1 or to the end of the prepositional phrase (**Figure 2A**), and then produced a continuation to complete the sentence (see Methods). Based on the continuations provided by the listeners at these two gating points, we can infer their online structural interpretations.

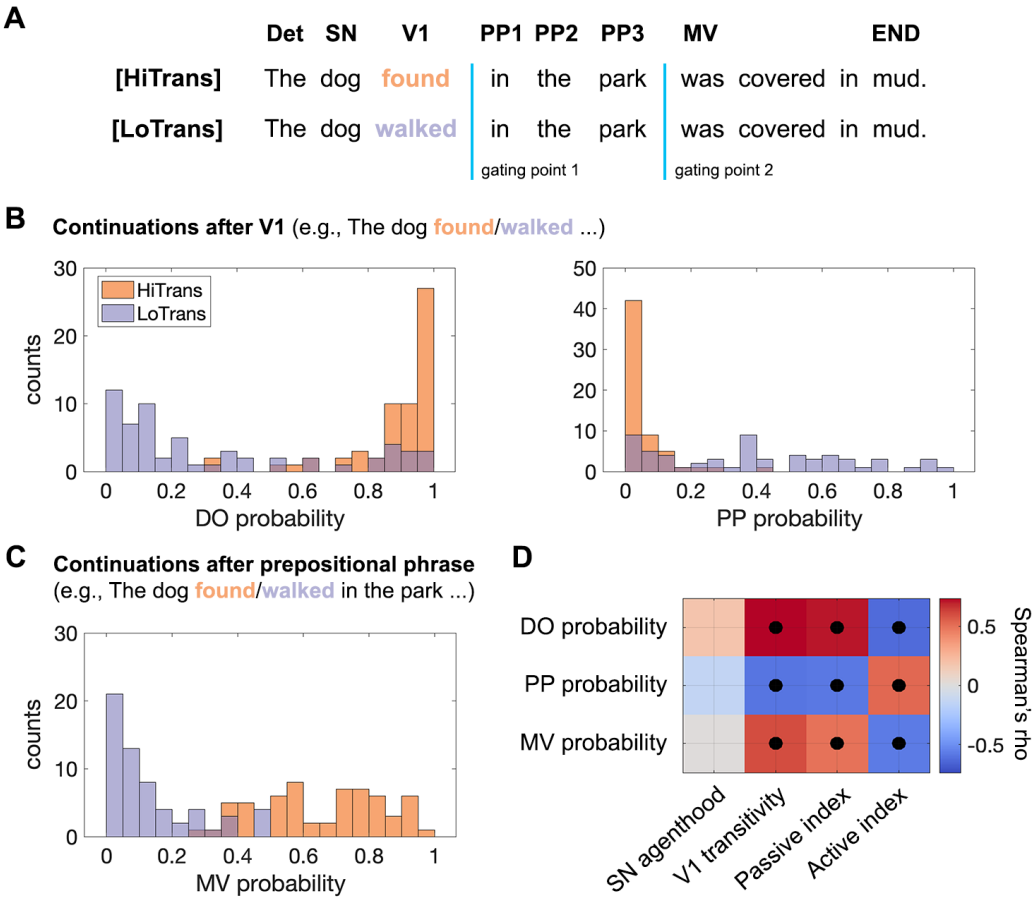

Figure 2. Human incremental structural interpretations derived from continuation pre-tests. (A) An example set of target sentences differing only in the transitivity of Verb1, HiTrans: high transitivity, LoTrans: low transitivity. Det: determiner, SN: subject noun, V1: Verb1, PP1-PP3: prepositional phrase, MV: main verb, END: the last word in the sentence. **(B)** Probability of a direct object (left) and a prepositional phrase (right) continuation after Verb1. **(C)** Probability of a main verb in the continuations after Verb1, which indicates an Active interpretation. **(D)** Correlations between corpus-based lexical constraints and probabilistic interpretations in the two pre-tests. (Spearman rank correlation, black dots indicate significance determined by 10,000 permutations, $P_{FDR} < 0.05$ corrected).

In the continuations after Verb1, a direct object was more likely to be found in HiTrans sentences, indicating a transitive use of Verb1, while an opposite pattern was found for a PP continuation, indicating an intransitive use of Verb1 (**Figure 2B**). As expected, the probability of a main verb (MV) in the continuations after the prepositional phrase was lower in LoTrans sentences (**Figure 2C**), suggesting that listeners preferred the Active interpretation and tended to interpret Verb1 as the main verb by the end of the prepositional phrase in LoTrans sentences, and vice versa in HiTrans sentences. Crucially, neither of the two pre-tests resulted in a

complete separation between HiTrans and LoTrans sentences; instead, they were characterized by two different but overlapping probabilistic distributions. This finding suggests that Passive and Active interpretations varied in plausibility in each sentence type before the actual main verb was presented, reflecting the probabilistic constraints jointly placed by the combination of the specific subject noun, Verb1, and the prepositional phrase in each sentence.

To relate these human interpretative preferences to the broader landscape of distributional language data, we developed corpus-based measures of the thematic role preference of the subject noun (i.e., how likely it is interpreted as an agent that conducts an action) and the transitivity of Verb1 in each sentence, from which we derived a Passive index and an Active index (see Methods). These indices separately capture the interpretative coherence between these two lexical properties towards Passive and Active interpretations. Both high subject noun agenthood and low Verb1 transitivity coherently preferred an Active interpretation as the prepositional phrase was heard (i.e., a high Active index), and vice versa for the Passive interpretation (i.e., a high Passive index). In accord with the *constraint-based* hypothesis, we found that human interpretative preference for the two types of sentences was significantly correlated with the lexical constraints placed by the subject noun and Verb1 (**Figure 2D**).

Incremental structural representations extracted from BERT

Next, we extracted structural representations at various positions in the same sentences from BERT and evaluated them according to the *constraint-based* hypothesis and human behavioural results. This evaluation is needed to motivate the use of BERT structural measures to reveal how the structured interpretation of a spoken sentence is incrementally built in the brain.

Typically, the structure of a sentence can be represented by a dependency parse tree (de Marneffe et al. 2006) where words are situated at different depths given their structural dependency (**Figure 3A**). Each edge links two structurally proximate words as being the head and the dependent separately (e.g., a verb and its direct object). However, such a parse tree is context-free, that is, it only captures the syntactic relation between each pair of words and abstracts away from the specific lexical (and higher order) contents of the sentence that constrain its structural interpretation. This context-free parse depth is always the same for words at the same position in sentences with the same structured interpretation (e.g., *"found"* and *"walked"* in either of the two parse trees in **Figure 3A**).

To obtain structural measures that also encode the specific lexical contents in a sentence, we adopted a structural probing technique (Hewitt and Manning 2019) to reconstruct a sentence's structure by estimating each word's parse depth based on their contextualized representations generated by BERT (see Methods). Note that BERT is a multi-layer DLM (24 layers in the version used in this study) which may distribute different aspects of its computational solutions over multiple layers. Accordingly, we trained a structural probing model for each layer, and selected the one with the most accurate structural representations while also including its neighboring layers to cover relevant upstream and downstream information. Following this

strategy, we used the BERT structural measures obtained from layers 12-16 with the best performance achieved in layer 14 (see **Appendix 1-figure 1** and Methods).

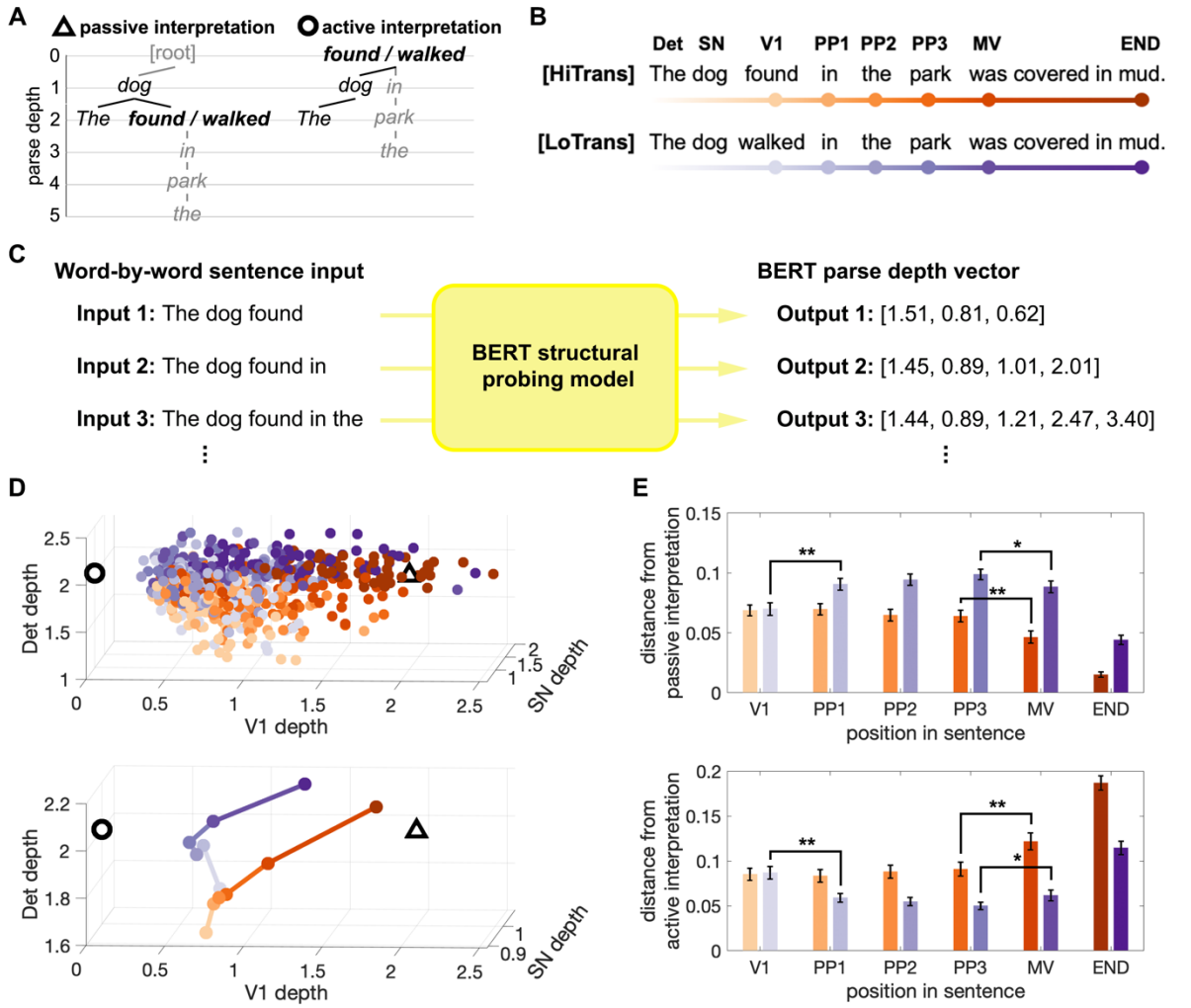

Figure 3. Incremental interpretation of sentential structure by BERT. (A) Context-free dependency parse trees of two plausible structural interpretations. Left: Passive interpretation where V1 is the head of a reduced relative clause. Right: Active interpretation where V1 is the main verb. **(B)** Incremental input to BERT structural probing model, with the lightness of dots encoding different positions in the target sentences. Det: determiner, SN: subject noun, V1: Verb1, PP1-PP3: prepositional phrase, MV: main verb, END: the last word in the sentence. **(C)** BERT structural probing model is trained to output a parse depth vector, representing the parse depths of all the words in the sentence input. The BERT parse depth for a specific word is updated incrementally as the sentence unfolds word-by-word. In this example, the parse depth of "*found*" increases with the presence of the prepositional phrase, indicating an increased preference for the Passive interpretation [according to the context-free parse depths in (A)]. **(D)** Incremental interpretation of the dependency between SN and V1 in the model space consisting of the parse depth of Det, SN and V1. Upper: Each colored circle represents the parse depth vector up to V1 derived at a certain position in the sentence [with the same color scheme as in (B)]. The hollow triangle and circle represent the context-free dependency parse vectors for Passive and Active interpretations in (A). Lower: incremental interpretation of the two target sentence types represented by the trajectories of median parse depth. **(E)** Distance from Passive and Active landmarks in the model space as the sentence unfolds [between each colored circle and the two landmarks in the upper panel of (D)] (two-tailed two-sample t-test, *: $P < 0.05$, **: $P < 0.001$, error bars represent SEM).

We input each sentence word-by-word to the trained BERT structural probing models, focusing on the incremental structural representation being built as it progressed from Verb1 to the main verb (see examples in **Figures 3B and 3C**). Note that we defined the first word after the prepositional phrase as the main verb since its appearance is sufficient to resolve the intended structure where Verb1 is a passive verb. We found that, for each type of sentences, the BERT parse depth of words at the same position formed a distribution ranging around the corresponding context-free parse depths in either the Passive or the Active interpretation (see **Appendix 1-figure 2**), suggesting a word-specific rather than position-specific structural representation. In addition, we quantified the contributions of words at different positions to the variances encoded in BERT parse depth vectors. Our analysis revealed that content words contributed significantly more than function words (i.e., the determiners) (see **Appendix 1-figure 3** for details).

Then we visualized BERT's word-by-word structural measures, focusing on the dependency between the subject noun and Verb1 that is core to the current interpretation of the sentence – whether the subject noun is the agent or the patient of Verb1. To this end, we built a 3-dimensional vector including BERT parse depths of the first three words up to Verb1 for each sentence (e.g., *"The dog found..."*). This 3D vector was updated incrementally with each additional word in the input, thereby capturing the dynamic interpretation of the structural dependency between the subject noun and Verb1, influenced by the context provided by subsequent words in a specific sentence. Like the probabilistic interpretation found within each type of sentences in human listeners, trajectories of individual HiTrans and LoTrans sentences are considerably distributed and intertwined (see the upper panel of **Figure 3D**), implying that BERT structural interpretations are sensitive to the idiosyncratic contents in each sentence.

To make sense of these trajectories, we also vectorized the context-free parse depth of the first three words indicating Passive and Active interpretations (**Figure 3A**) separately and located them in the 3D vector space as landmarks (hollow triangle and circle in **Figure 3D**), so that the plausibility of either interpretation can be estimated by a sentence's distance from its landmark. As shown by the trajectories of the median BERT parse depth of the two sentence types (see the lower panel of **Figure 3D**), in general, HiTrans sentences continuously moved towards the Passive interpretation landmark after Verb1, with a significant change of distances detected at the main verb (see the orange bars in **Figure 3E**). LoTrans sentences started by approaching the Active interpretation landmark but were reorientated to the Passive counterpart with the appearance of the actual main verb, with significant changes of distances detected at both Verb1 and main verb (see the purple bars in **Figure 3E**). These results resemble the pattern of human interpretative preference observed in the continuation pre-tests, where the Passive and Active interpretations were separately preferred in HiTrans and LoTrans sentences by the end of the prepositional phrase in a probabilistic manner (**Figure 2C**), before the Passive interpretation was established with the appearance of the actual main verb.

BERT structural measures are correlated with constraints driving human interpretation

To further assess whether BERT's preferences for structural interpretation align with the constraints considered by human listeners during speech comprehension, we correlated BERT

structural measures with relevant corpus-based measures and human behavioural data (see Methods).

We first focused on BERT's interpretative mismatch quantified as the distance between an unfolding sentence and each of the two landmarks in the model space, which was dynamically updated as the sentence unfolded (**Figure 3D**). Consistently, from the incoming prepositional phrase to the main verb, sentences that are closer to the Passive landmark in the model space have higher Verb1 transitivity, a higher Passive index but a lower Active index, while sentences closer to the Active interpretation landmark exhibited higher Active index and lower Passive index (**Figures 4A** and **4B**). Moreover, at the beginning of the prepositional phrase, the change of distance towards either interpretation landmark between two consecutive words is also correlated with these constraints (**Figures 4C** and **4D**), suggesting an immediate update in the structural interpretation in combination with the accumulated constraints from the preceding subject noun and Verb1.

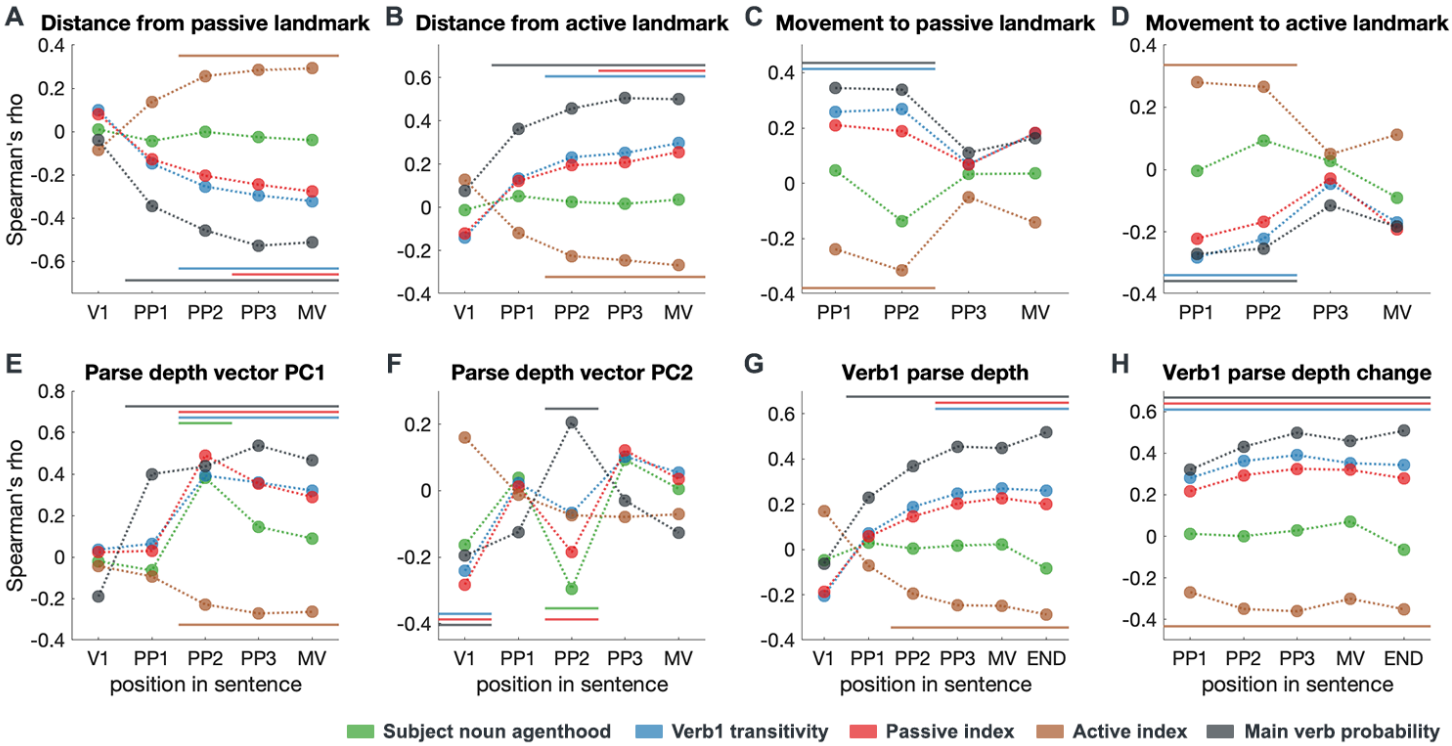

Figure 4. Correlation between incremental BERT structural measures and explanatory variables. BERT structural measures include **(A**, **B)** BERT interpretative mismatch represented by each sentence's distance from the two landmarks in model space (Figure 3D); **(C**, **D)** Dynamic updates of BERT interpretative mismatch represented by each sentence's movement to the two landmarks; **(E**, **F)** Overall structural representations captured by the first two principal components (i.e., PC1 and PC2) of BERT parse depth vectors; **(G**, **H)** BERT Verb1 (V1) parse depth and its dynamic updates. Explanatory variables include lexical constraints derived from massive corpora and the main verb probability derived from the continuation pre-test (Spearman correlation, permutation test, $P_{FDR} < 0.05$, multiple comparisons corrected for all BERT layers, results shown here are based on layer 14, see **Appendix 1-figures 4-6** for the results of all layers, see **Appendix 1-figure 7** for the dynamic change of Verb1 parse depth); PP1-PP3: prepositional phrase, MV: main verb, END: the last word in the sentence.

Moreover, we found that both the incremental BERT parse depth vectors as a whole (which are captured by their principal components) and the BERT parse depth of Verb1 (which is the most indicative marker of the interpretation preferred) are correlated with the constraints placed by the subject noun and Verb1 (**Figures 4E** to **4H**). Moreover, the significant effects consistently found as the sentence unfolds suggest that properties of preceding words are used to constrain the interpretation of the upcoming input, which is key to resolving discontinuous structural dependencies. In addition, we found that BERT structural interpretations were also correlated with the main verb probability in the continuation pre-test which directly reflects human interpretation preference (black bars in **Figure 4**).

Overall, these results illustrated, at which position in a sentence, relevant lexical constraints started being encoded by BERT, which also validated the contextualized BERT structural measures in terms of the *constraint-based* hypothesis and human behavioural results, and motivated the use of them to probe the neural processes involved during the incremental structural interpretation of spoken sentences.

Neural dynamics of incremental structural interpretation

To study how the structured interpretation of a spoken sentence is built word-by-word in the brain, we used ssRSA to test the incremental BERT structural measures in source-localized EMEG collected when the same sentences were delivered to human listeners (see **Figure 5** for the pipeline of ssRSA). This combination of methods gains improved neurocomputational specificity by probing the spatiotemporally resolved neural activity with detailed structural representations rather than the entire hidden states of BERT. We compared the representational geometry of BERT structural measures with that of neural responses inside a spatiotemporal searchlight moving across the cortical surface, significant model fits showed when and where the incremental structural interpretations or relevant lexical constraints emerge and update in the brain. Given the probabilistic interpretations in BERT and human listeners reported above, we combined HiTrans and LoTrans sentences as one group to increase the range of pair-wise dissimilarity to be modelled in ssRSA.

We began with the BERT parse depth vector containing the parse depth of each word in an incremental input, providing a dynamic structural representation updated as the sentence unfolded. Then, we tested the interpretative mismatch between the incremental BERT parse depth vector and the corresponding context-free parse depth vector for the Passive or the Active interpretation. The degree of this mismatch is proportional to the evidence for or against the two interpretations, i.e., the smaller the distance, the more positively loaded this interpretation. Besides these two measures based on the entire incremental input, we also focused on Verb1 since the potential structural ambiguity lies in whether Verb1 is interpreted as a passive verb or the main verb. Given the context-free parse depth of Verb1 that is 2 in the passive interpretation and 0 in the Active interpretation (**Figure 3A**), with each incoming later word, an increased BERT Verb1 parse depth towards 2 or a decreased value towards 0 reflects separately the preference biased to a Passive or an Active interpretation (**Appendix 1-figure 7**). All quantitative measures tested in ssRSA are summarized in **Appendix 1-table 1**).

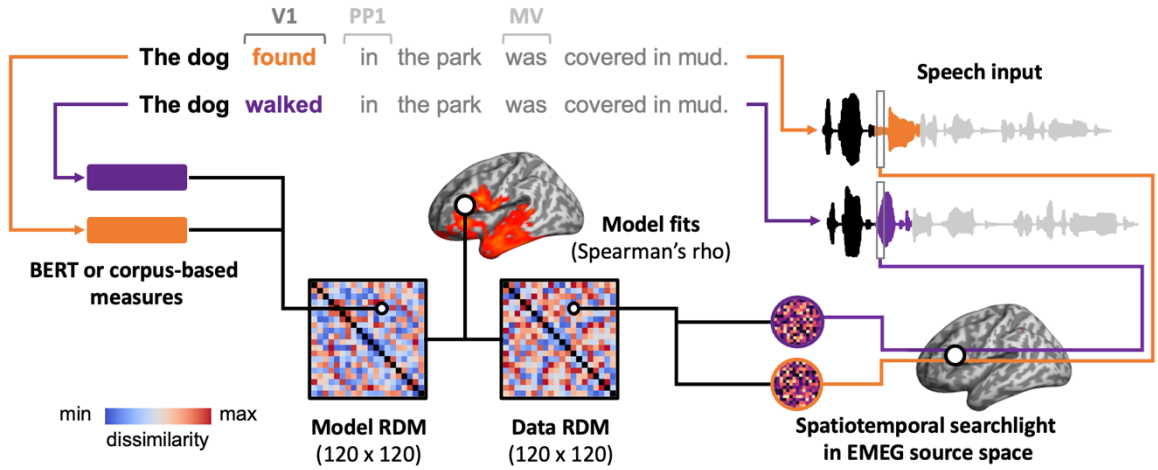

Figure 5. Illustration of the pipeline for ssRSA. For each pair of sentences, we extract their BERT or corpus-based measures and calculate the dissimilarity between these measures, resulting in a model representational dissimilarity matrix (RDM). Meanwhile, we also extract the neural activity recorded while participants are listening to these sentences and calculate their dissimilarity to create a data RDM. Specifically, we use a spatiotemporal searchlight in EMEG source space, which moves across the brain and captures the neural activity within a 10-mm-radius sphere over a 60-ms sliding time window. By correlating the model RDM with data RDMs from all spatiotemporal searchlights, we can identify whether, and if so, when and where the brain represents the information captured by the model RDM. The ssRSA is conducted in V1, PP1 and MV epochs respectively, with HiTrans and LoTrans sentences combined as one group (i.e., 120 sentences in total).

For the listener's neural activity, we focused on three critical epochs in each sentence: (a) Verb1 – when its structural dependency with the preceding subject noun was initially established despite potential ambiguity, (b) the preposition – when the initial structural interpretation started being updated, to be either strengthened or weakened by the incoming preposition phrase, and (c) main verb – when the intended Passive interpretation was finally confirmed. We aligned the continuous EMEG data to the onset of Verb1, the preposition, and the main verb respectively and obtained three 600-ms epochs.

We found that the incremental BERT parse depth vectors exhibited significant fits to brain activity consistently in all three epochs, as the corresponding word was being heard at that time (**Figures 6A** to **6C**). In Verb1 epoch, effects in bilateral frontal and anterior-to-middle temporal regions started immediately from Verb1 onset and continued until the uniqueness point – the point at which the word has been uniquely identified. These early effects could be due to the different subject nouns included in the BERT parse depth vectors. While the BERT parse depth of Verb1 *per se* showed similar effects but with greater duration which peaked exactly at Verb1 uniqueness point (**Appendix 1-figure 10**). As the sentence unfolded, effects of BERT parse depth vectors were found in the left fronto-temporal regions in the two later epochs, starting after the recognition of the preposition or the main verb separately.

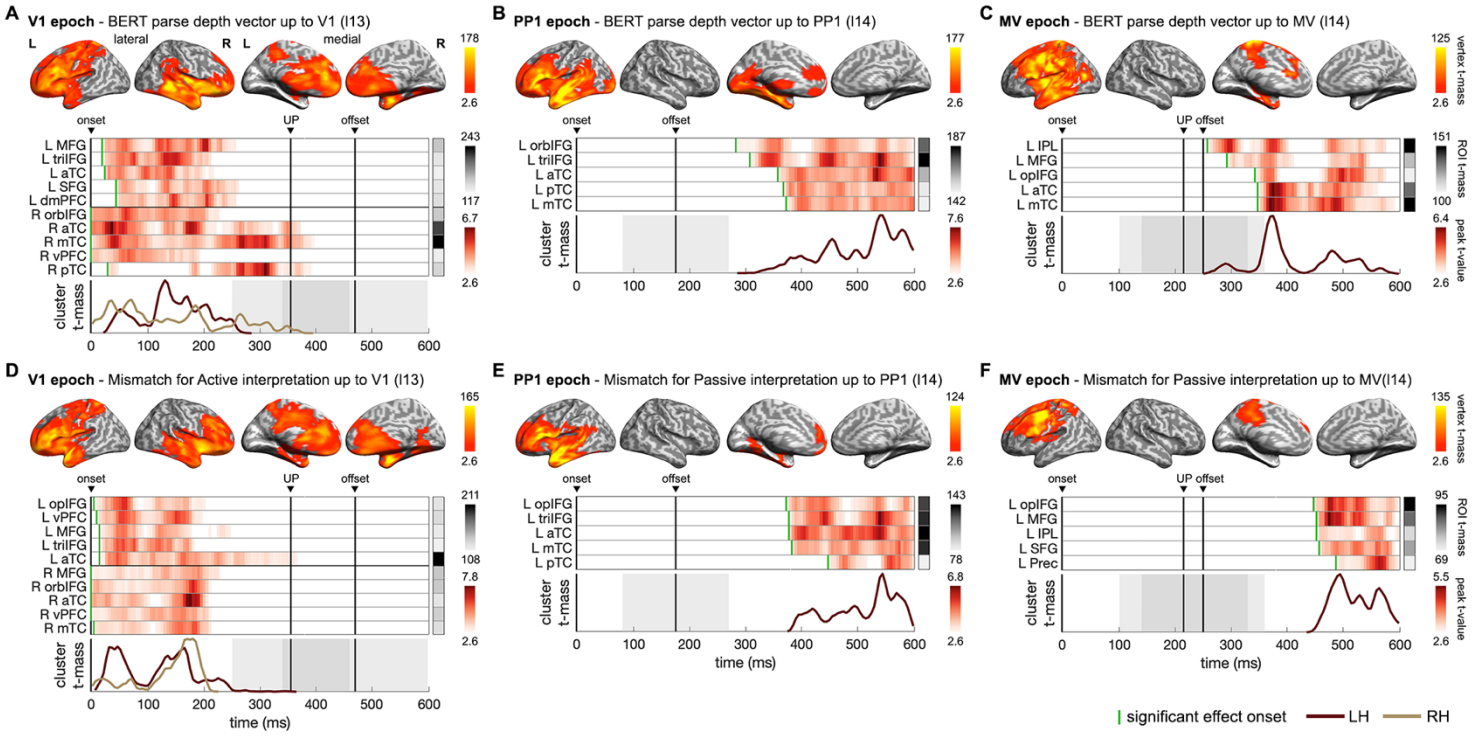

Figure 6. Neural dynamics underpinning the emerging structure and interpretation of an unfolding sentence. (A-C) ssRSA results of BERT parse depth vector up to Verb1 (V1), the preposition (PP1) and the main verb (MV) in epochs separately time-locked to their onsets. **(D-F)** ssRSA results of the mismatch for the preferred structural interpretation (the specific BERT layer from which BERT structural measures were derived was denoted in parentheses). From top to bottom in each panel: vertex t-mass (each vertex's summed t-value during its significant period); heatmap of time-series of ROI peak t-value (the highest t-value in an ROI at each time-point) with a green bar indicating effect onset and ROI t-mass (each ROI's summed mean t-value during its significant period); cluster t-mass time-series (summed t-value of all the significant vertices of a cluster at each time-point). [cluster-based permutation test, vertex-wise $P < 0.01$, cluster-wise $P < 0.05$ in (A-E); marginal significance in (F) with cluster-wise $P = 0.06$]. Solid vertical lines indicate the timings of onset, average uniqueness point (UP), and average offset of the word time-locked in the epoch with grey shades indicating the range of one SD. LH/RH: left/right hemisphere. See **Appendix 1-table 2** for full anatomical labels. See **Appendix 1-figure 8** for Spearman's rho time-series of ROIs in individual participants, see **Appendix 1-figure 9** for the significant results of other BERT layers in the MV epoch.

Turning to the interpretative mismatch for the two possible interpretations, we only observed significant effects of the mismatch for Active interpretation in Verb1 epoch (**Figure 6D**). However, it was the mismatch for Passive interpretation that fitted brain activity in the preposition and main verb epochs (**Figures 6E** and **6F**, marginal significance in main verb epoch with cluster-wise $P = 0.06$). These results suggest that listeners, in general, tended to have an initial preference for an Active interpretation (even before the recognition of Verb1) but might start favoring a Passive interpretation when the prepositional phrase began to be heard. This finding is consistent with the tendency to process the first noun encountered at the beginning of a sentence as the agent (Bever 1970; Jackendoff and Jackendoff 2002; Mahowald et al. 2023). Note that our approach does not constitute a direct test for the hypothesis of parallel parsing, as we did not uncover evidence supporting the maintenance of parallel representations

of different syntactic structures in the brain; rather, we only found one preferred structure in each epoch.

Effects of the BERT parse depth vectors and those of the interpretative mismatch for the preferred structural interpretation have substantial overlaps in terms of their spatiotemporal patterns in the brain, characterized primarily by a transition from bilateral to left-lateralized fronto-temporal regions as the sentence unfolds. Across the three epochs, the most sustained effects were observed in the left inferior frontal gyrus (IFG) and the anterior temporal lobe (ATL). Notably, with the identification of the actual main verb, effects of the eventually resolved structure also involved regions in the left prefrontal and inferior parietal regions (**Figure 6C**) which belong to the multiple-demand network (Duncan 2010). The involvement of the prefrontal regions could be indicative of the varying working memory demands (e.g., the different number of open nodes in the sentence structures for Active and Passive interpretations before the actual main verb is recognized) for building the structure of the unfolding sentence (Pallier *et al.* 2011; Nelson *et al.* 2017).

Structural ambiguity resolution probed using BERT Verb1 parse depth

As mentioned above, the potential ambiguity between a Passive and an Active interpretation centers around whether Verb1 is considered as a passive verb or the main verb, which is resolved upon the appearance of the actual main verb. We probed how this is implemented in the brain using the dynamic BERT parse depth of Verb1. Specifically, the cognitive demands required by this resolution process can be characterized by the change between the updated BERT parse depth of Verb1 when the actual main verb is presented and its initial value when Verb1 is first encountered (see **Appendix 1-figure 7** for the dynamic change of BERT Verb1 parse depth).

We first tested the change of Verb1 parse depth in the main verb epoch. Significant fits to brain activity emerged in the left posterior temporal and inferior parietal regions upon the main verb uniqueness point, and then extended to more anterior temporal regions (**Figure 7A**). After the main verb offset, the declining effects of the Verb1 parse depth change in the left anterior temporal region seamlessly overlapped with the arising effects of the updated Verb1 parse depth (**Figures 7B** and **7C**). These results indicate that the recognition of the actual main verb immediately triggered an update of the previous interpretation of Verb1, with the resolved interpretation emerged in the left temporal lobe and was later delivered to the right posterior temporal and parietal areas. It is also worth noting that the left hippocampus was activated for both measures of Verb1 parse depth after the actual main verb is recognized, suggesting that the episodic memory of experienced events might contribute to the updating of structural interpretations (Bicknell et al. 2010; Metusalem et al. 2012; Altmann and Ekves 2019). These results address the dynamic update of structured interpretation by focusing on the BERT parse depth of Verb1, which complements those of the interpretative mismatch based on the incremental BERT parse depth vector incorporating constraints of all the words heard so far (**Figure 6F**).

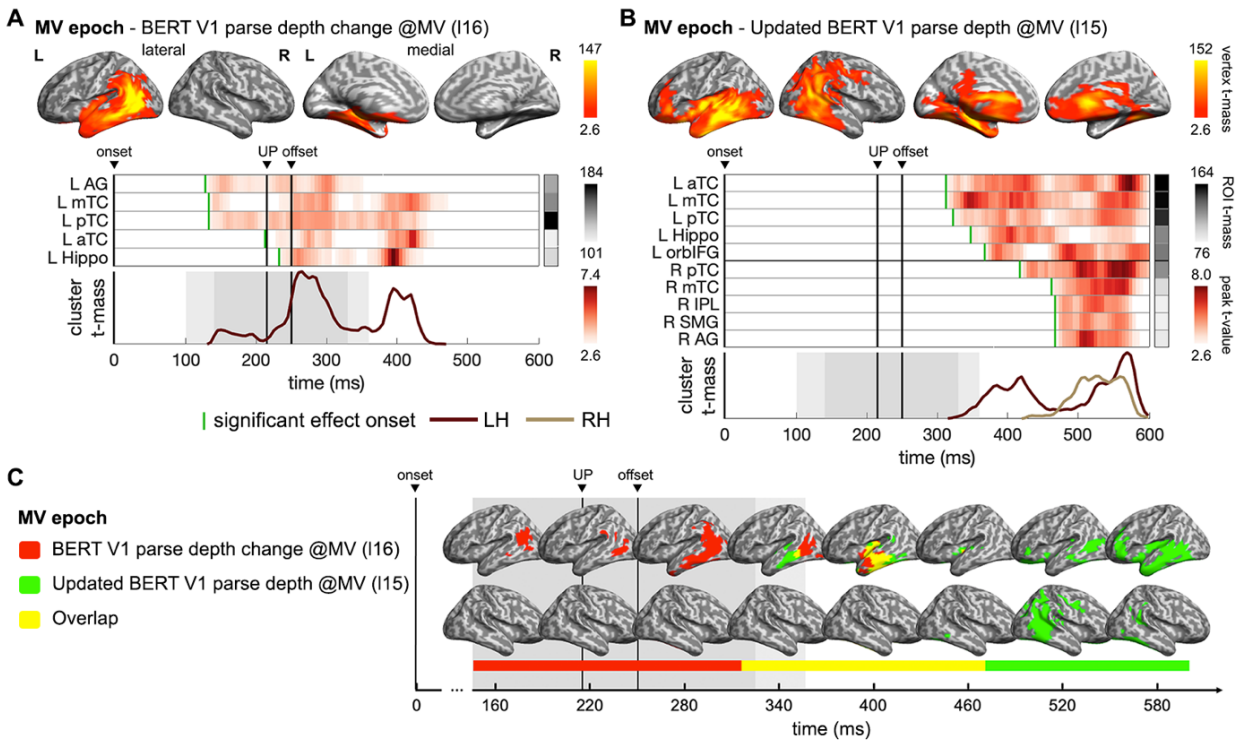

Figure 7. Neural dynamics updating the incremental structural interpretation. (A) ssRSA results of BERT Verb1 (V1) parse depth change at the main verb (MV) relative to the parse depth of V1 when it is first encountered. **(B)** ssRSA results of the updated BERT V1 parse depth when the input sentence reaches the MV. **(C)** Spatiotemporal overlap between the effects in (A) and (B). (cluster-based permutation test, vertex-wise $P < 0.01$, cluster-wise $P < 0.05$). See **Appendix 1-figure 14** for Spearman's rho time-series of ROIs in individual participants.

Emergent structural interpretations driven by multifaceted constraints in the brain

Next, we further asked how the multifaceted constraints, considered by human listeners and encoded in BERT parse depths, drive the structured interpretation in the brain. When and where in the brain do these constraints emerge? How are their neural effects related to those of the final resolved sentential structure? To address these questions, we first tested the subject noun thematic role properties obtained from corpus data. Significant effects of agenthood and patienthood were found in the preposition epoch (**Figure 8A**) and in the main verb epoch (**Figure 8B**) separately. Notably, effects of the subject noun itself preceded those of incremental BERT parse depth vectors modelling the sentence fragments in the same epoch (compare **Figure 8A** with **Figure 6B**, and **Figure 8B** with **Figure 6C**). These findings indicate that subject noun thematic role might be evaluated before building the overall structural interpretation of the utterance delivered so far. Specifically, the initial preference for an Active interpretation during Verb1, while present as the preposition started (i.e., subject noun agenthood in **Figure 8A**), was superseded by the preference for a Passive interpretation as the rest of the prepositional phrase (**Figure 6A**) and the main verb (**Figure 6B**) were heard.

Despite being jointly constrained by subject noun thematic role preference and Verb1 transitivity in a probabilistic manner, the structural interpretation temporarily held just before the recognition of the actual main verb could differ across sentences (e.g., Passive interpretation in "*The dog found in the park...*" and Active interpretation in "*The dog walked*

in the park...”). Therefore, in contrast to the Passive or Active index specialized for one particular structural interpretation, we constructed a non-directional index that merely quantifies the degree of interpretative coherence for one interpretation, whether Passive or Active (see Methods for details). Thus, a higher value only indicates greater interpretative coherence between the subject noun and Verb1 regardless of which interpretation is preferred.

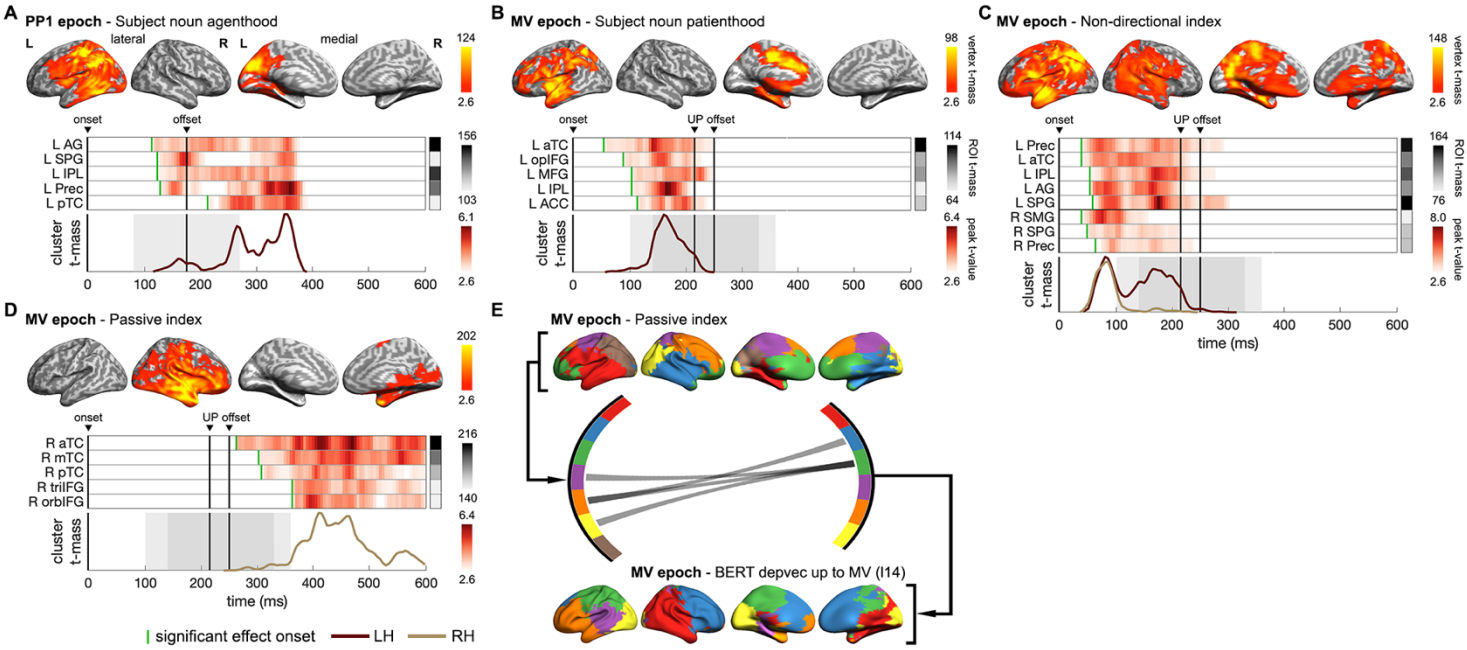

Figure 8. Neural dynamics of multifaceted probabilistic constraints underpinning incremental structural interpretations. (A, B) ssRSA results of SN agenthood and SN patienthood (i.e., plausibility of SN being the agent or the patient of V1) in PP1 and MV epochs separately. **(C)** ssRSA results of non-directional index (i.e., interpretative coherence between SN and V1 regardless of the structure preferred) in MV epoch. **(D)** ssRSA results of Passive index (i.e., interpretative coherence for the Passive interpretation) in MV epoch. **(E)** Influence of the Passive interpretative coherence on the emerging sentential structure in MV epoch revealed by the Granger causal analysis (GCA) based on the non-negative matrix factorization (NMF) components of whole-brain ssRSA results (see **Appendix 1-figure 16** for more details) [(A-D) cluster-based permutation test, vertex-wise $P < 0.01$, cluster-wise $P < 0.05$; (E) permutation test $P_{FDR} < 0.05$]. See **Appendix 1-figure 15** for Spearman’s rho time-series of ROIs in individual participants.

Effects of this non-directional measure of interpretative coherence appeared very soon after the main verb onset in both hemispheres and lasted till its offset (**Figure 8C**), suggesting an immediate evaluation of the previously integrated constraints from the subject noun and Verb1 when a listener realized that the sentence had not finished yet. Moreover, these effects roughly co-occurred with the effects of subject noun patienthood (compare **Figures 8B** and **8C**), indicating that a patient role for the subject noun was considered as the main verb was being recognized. Intriguingly, the most sustained regions associated with this non-directional index, including the left ATL, angular gyrus (AG) and precuneus, are also the classical areas of the default mode network (DMN). This finding is consistent with recent claims that the DMN integrates external input with internal prior knowledge to make sense of an input stimulus such as speech (Yeshurun *et al.* 2021). In particular, precuneus and AG have been found to be involved in building thematic relationships and event structures (Baldassano et al. 2017; Humphreys et al. 2021).

Following the declining effects of the non-directional index upon the recognition of the main verb, we found significant effects of the Passive index in right anterior fronto-temporal regions (**Figure 8D**), suggesting that the intended Passive interpretation was eventually established in all sentences. Previous studies have revealed that the relatively narrow sentence-specific information and the broad world knowledge are processed in the left and right hemispheres separately (Jung-Beeman 2005; Metusalem et al. 2016; Troyer et al. 2022). Relevant to this, in the main verb epoch, we found effects of the BERT parse depth vector and those of the Passive index in the left and right hemispheres respectively, arising almost at the same time as the main verb was recognized (compare **Figure 6C** and **Figure 8D**). Therefore, a critical question is whether and how the online structural interpretation of a specific sentence is facilitated by the interpretative coherence conjured up from lexical constraints that also depend on broad world knowledge (e.g., thematic role).

To address this question, we adopted non-negative matrix factorization to decompose the whole-brain ssRSA fits of the Passive index and the BERT parse depth vector found in the main verb epoch into two sets of components given their temporal synchronizations (see Methods). We then conducted multivariate Granger causality analyses (GCA) to infer directed connections among them. We found only GC connections from the components of Passive index to those of BERT parse depth vector (**Figure 8E**). Specifically, we identified information flows from the right hemisphere components of the Passive index to the left hemisphere components of BERT parse depth vector, suggesting that a sentence's structure represented in the left hemisphere might be influenced by the coarse estimate of the event plausibility concurrently determined by broad world knowledge in the right hemisphere (Jung-Beeman 2005) (see **Appendix 1-figure 16** for more details).

Comparisons between BERT and corpus/behavioural measures in fitting neural activity

To directly assess the performance of BERT structural measures with that of traditional measures extracted from corpus or behavioural data in fitting listeners' neural activity, we also conducted ssRSA with model RDMs of corpus-based or behavioural measures. In the Verb1 epoch, we tested Verb1 transitivity obtained from either corpus data or human continuations, however, neither of them exhibited significant model fits, which contrasted with the pronounced effects of BERT Verb1 parse depth (**Appendix 1-figure 11**). Similarly, in the PP1 and MV epochs, the probabilities of PP and MV continuations, as determined from behavioral data, did not show any significant model fits (**Appendix 1-figure 12** and **Appendix 1-figure 13**). Furthermore, the effects of BERT parse depth vector in these two epochs (**Figure 6B** and **6C**) remained largely unchanged after controlling for the variance explained by the behavioural measures. These findings suggest that BERT structural measures, compared to corpus-based and behavioral measures, are better at fitting the neural dynamics during incremental speech comprehension. This might be attributed to the capacity of DLMs to capture more nuanced and contextually rich representations (Linzen and Baroni 2021; Pavlick 2022).

Discussion

In this study, we investigated the neural dynamics involved in constructing structured interpretations from speech. We combined spatiotemporally resolved brain activity of human listeners, quantitative structural representations derived from a DLM (i.e., BERT), and corpus-based and behavioural measures. Our study revealed the emergence and update of a structured interpretation, jointly constrained by different lexical properties related to both linguistic and non-linguistic world knowledge, in an extensive set of brain regions beyond the core fronto-temporal language network. Specifically, our results show (1) a shift from bi-hemispheric lateral frontal-temporal regions to left-lateralized regions in representing the current structured interpretation as a sentence unfolds, (2) a pattern of sequential activations in the left lateral temporal regions, updating the structured interpretation as syntactic ambiguity is resolved, and (3) the influence of lexical interpretative coherence activated in the right hemisphere over the resolved sentence structure represented in the left hemisphere. These findings provide empirical evidence for the *constraint-based* approach to sentence processing and deepen the understanding of specific spatiotemporal patterning and neuro-computational properties underpinning incremental speech comprehension.

Using artificial neural networks (ANNs) to study the neural substrates of human cognition complements the long-time pursuit of generative rules and interpretable models (Kriegeskorte and Douglas 2018). ANNs have informed our understanding of various cognitive processes in the brain by providing quantifiable predictions that aim to connect behaviors and relevant neural activity (Yamins and DiCarlo 2016; Rabovsky et al. 2018; Donhauser and Baillet 2019; Kietzmann et al. 2019; Yang et al. 2019; Bao et al. 2020; Sheahan et al. 2021; Doerig et al. 2023; Giordano et al. 2023). This is crucial for quantifying the outcome of complex, interrelated constraints that arise in specific contexts, such as spoken sentences, and constructing the representational geometry to be probed in the brain. Where DLMs are concerned, recent studies have systematically compared the internal representations of DLMs to those observed in the human brain during language processing, which highlights the importance of predictive coding and contextual information (Schrimpf *et al.* 2021; Caucheteux *et al.* 2022; Caucheteux and King 2022; Goldstein *et al.* 2022; Heilbron *et al.* 2022; Toneva *et al.* 2022; Caucheteux *et al.* 2023). Furthermore, these studies have motivated the use of DLMs as a computational tool, or hypothesis, to study the neural substrates of language.

Here we asked a more specific question, that is, how a sequence of spoken words is incrementally structured and coherently interpreted in the brain. Our goal was to use quantitative measures of sentence structure that capture the interplay between different types of constraints that simultaneously influence this process. As a potential solution, we extracted detailed structural measures specific to the contents in each sentence from the hidden states of BERT, which was trained on massive corpora from real-life language use. Although DLMs such as BERT are not specifically designed to parse sentences, they can learn from training corpora the multi-dimensional properties related to sentence structure and dependency (Manning *et al.* 2020). In line with this, our analyses confirmed that BERT structural measures

incorporate relevant lexical constraints and that they exhibit both behavioural and neural alignments with human listeners.

Taking advantage of the contextualized BERT structural measures, our ssRSA results provide neural evidence for the construction of a coherent interpretation driven by the interaction between linguistic and non-linguistic knowledge evoked by individual words as they are heard sequentially in a spoken sentence. Specifically, neural representations of an unfolding sentence's structure initially emerged in bilateral fronto-temporal regions and became left-lateralized when more complex syntactic properties, rather than canonical linear adjacency, were considered to build a structured interpretation (e.g., beyond Verb1 in our stimulus sentences). Meanwhile, we found right-hemisphere activations associated with broad world knowledge, which is essential for understanding the intended meaning conveyed by the speaker (Bicknell *et al.* 2010). In addition to the core fronto-temporal language network, we found that the multiple-demand network and the default mode network were also involved during online construction of structured interpretations, which may reflect additional cognitive demands for resolving potential structural ambiguity and evaluating the plausibility of underlying events (Smallwood et al. 2021).

Moreover, our results show that, compared to corpus-based and behavioural measures, BERT structural measures are more effective in fitting listeners' neural activity, possibly due to its advanced ability in modelling specific contexts within each sentence. Nevertheless, it is important to recognize the important role of corpus-based and behavioral measures as explanatory variables. They are crucial not only in interpreting BERT measures but also in understanding their alignment with listeners' neural activity. This includes, for instance, the temporal sequence of activations of key lexical constraints and the emerging structure of a sentence (e.g., effects of subject noun patienthood leading those of BERT parse depth vector in the MV epoch, as seen in **Figure 8B** and **Figure 6C**) and the spatial distribution of their model fits in the brain (e.g., contrasting model fits of Passive index and BERT parse depth vector in the MV epoch across different hemispheres, as shown in **Figure 8D** and **Figure 6C**). Such an integrative approach allows for a more comprehensive understanding of the complex mental processes underpinning speech comprehension, which takes advantages of the interpretability of traditional measures and the deep contextualized representations of DLMs.

There are two points to note about the use of BERT. Firstly, unlike autoregressive DLMs trained using left-to-right attention and the next-word prediction task, BERT is trained to predict masked words in a sentence with a bi-directional attention mechanism. The additional right-to-left attention provides updated representations of preceding words every time an incoming word is added to the input (e.g., representation of *"dog"* in *"The dog…"* is different from that in *"The dog found…"*). This feature of BERT is useful for tracking the dynamic change of the representation of a specific word as its context evolves (e.g., Verb1 in this study), particularly in sentences with structural ambiguity. Although autoregressive DLMs also update hidden states as the input unfolds and could be used to study complex sentential structures (Jurayj et al. 2022), the updated contextual effects are reflected in the hidden states of the right-most incoming word, while those of the preceding words on the left remain unchanged (i.e.,

the representation of *"dog"* is the same in *"The dog…"* and *"The dog found…"*). This is different from BERT, where the updated contextual effects are reflected in the hidden states of all preceding words.

Secondly, although we input each sentence word-by-word to BERT, however, unlike human listeners or recurrent neural networks, BERT process two consecutive inputs (e.g., *"The dog…"* and *"The dog found…"*) independently, and there is no direct relationship between the hidden states of these two inputs. In fact, human listeners would not start over from the beginning of a sentence as it unfolds word-by-word, but update it continually as each word is heard and use whatever information currently available to build a coherent interpretation (McRae and Matsuki 2013). This difference, however, does not impede our objective of extracting contextualized structural representations at critical points within a sentence. In the case of BERT case, the representation of each word is continuously updated in a bi-directional manner as a new word is added. This process accounts for the constraints imposed by all the words of the input and their interactions, forming a coherent interpretation. Nonetheless, for other hypotheses in speech comprehension, such as parallel parsing and how various grammatically correct sentence structures are maintained and compete in the brain, DLMs with recurrent memory might be more suitable. Such models can better stimulate the continuous, dynamic updating of interpretations that characterizes human sentence processing.

In summary, recent developments in DLMs have shown great potential in capturing the dynamic interplay between syntax, semantics, and broader world knowledge that is essential for successful language comprehension. The empirical evidence from this study supports the notion that DLMs, when utilized as potential models of brain computation and integrated with advanced neuroimaging techniques within a well-defined framework can offer significant insights in to human cognition (Kriegeskorte and Douglas 2018; Doerig *et al.* 2023). Future DLMs, especially those with more human-like model architecture (McClelland et al. 2020) and subjected to rigorous evaluation (Binz and Schulz 2023), hold the potential to shed light on the neural implementation of various rapid incremental processes that support the rapid transition from sound to meaning in the brain.

Materials and Methods

Participants

Seventeen right-handed native British English speakers participated in this study and provided written consent. One participant was excluded from subsequent analysis due to sleepiness during EMEG scanning, the other sixteen participants were included in the following analyses (aged between 19 and 38 years, 26.5 years on average; 7 females). All participants had normal hearing, and none had any pre-existing neurological condition or mental health issues. This study was approved by the Cambridge Psychology Research Ethics Committee.

Stimuli

We constructed 60 sets of six spoken sentences (360 in total) with varying sentential structures. As shown in the example sentence set (**Appendix 1-table 3**), unambiguous (UNA), high transitivity (HiTrans) and low transitivity (LoTrans) sentences contain a long-distance dependency between the subject noun and the main verb introduced by a full or reduced relative clause inserted in between. Whereas there is no long-distance dependency in the sentences of passive (PAS) and two direct object (DO1 and DO2) conditions.

Unlike the first verb (Verb1) in the UNA sentences which is unambiguously interpreted as the head of a relative clause, Verb1 in both HiTrans and LoTrans sentences can also be considered alternatively as the "main verb" before the actual main verb (e.g., *was covered* in the example set in **Appendix 1-table 3**) was heard. By varying the nature of Verb1 in the reduced relative clause (e.g., *found/walked*), we manipulated the preference for the two plausible structural interpretations in HiTrans and LoTrans sentences before the appearance of the actual main verb (i.e., a Passive interpretation where Verb1 is the head of a relative clause - the subject noun undergoes the action specified by Verb1; an Active interpretation where Verb1 is the main verb - the subject noun performs the action specified by Verb1).

Specifically, in the LoTrans sentences, Verb1 was selected to be optionally transitive according to CELEX (Baayen et al. 1993) and Google n-gram corpus (books.google.com/ngrams), meaning that it can either take a direct object or not. Thus, a listener could be initially "garden-pathed" into the alternative Active interpretation where the Verb1 is considered as the main verb when a following prepositional phrase fits its intransitive use. Whereas Verb1 in HiTrans sentences was selected to have a higher preference for taking a direct object than Verb1 in LoTrans sentences [subcategorization frame (SCF) probability for direct object according to VALEX (Korhonen et al. 2006): HiTrans 0.71 ± 0.16, LoTrans 0.44 ± 0.19, two-tailed two-sample t-test, $t_{(117)} = 8.45$, $p = 9.3 \times 10^{-14}$]. Therefore, Verb1 in HiTrans sentences was more likely to be recognised as the head of a reduced relative clause given the appearance of a prepositional phrase (e.g., *in the park*) rather than a highly expected direct object.

In all the six types of sentences, the subject noun phrase comprised a single-word noun and a preceding determine "*The*". In each sentence set, sentences of the first four types had the same subject noun, while DO1 and DO2 sentences shared a different subject noun. The reduced relative clause in HiTrans and LoTrans sentences consisted of a head verb, i.e., Verb1 (e.g., *found/walked*) which was followed by a three-word prepositional phrase (e.g., *in the park*).

Note that, in 15 out of the 60 sentence sets, the actual main verb in UNA, RR and GP conditions was preceded by an auxiliary verb (e.g., "*was covered*" in the example set in **Appendix 1-table 3**). In the following analyses, we defined the first word after the prepositional phrase as the main verb since its appearance is sufficient to resolve the intended Passive interpretation where Verb1 is a passive verb (i.e., the head of a reduced relative clause).

Note that, although UNA, PAS, DO1 and DO2 conditions were not included in subsequent analyses, they added variety to the types of syntactic construction of the stimuli and ecological validity of the experiment, which also prevented potential adaption to a particular sentence structure.

Procedure

The experimental stimuli (360 spoken sentences recorded by a female native British English speaker with a neutral intonation throughout) were equally divided into four blocks with 90 experimental trials in each. To maintain participants' attention while they were listening to the stimuli, the experimental trials in each block were interspersed with 9 additional trials consisting of questions related to the contents in the preceding sentence. These questions were presented in written form on the screen, and a "yes" or "no" response was required by button pressing. Each of these question trials was followed by a filler trial (including a normal spoken sentence outside the experimental stimuli) to ensure that no residual task effects would be picked up in the next experimental trial. Each block started with two filler trials. All question trials and filler trials (20 in each block) were excluded from the following analyses. The order of blocks and the order of trials within each block were pseudorandomized across participants.

Each experimental trial began with a fixation cross presented at the centre of the screen with a random period ranging from 750ms to 1250ms (1000ms on average) before the onset of the spoken sentence. Participants were asked to look at the fixation cross and to avoid eye movement or blinking while listening to the spoken sentences. There was a 1000ms silence from the end of each sentence followed by a "blink cue" that lasted for 1400ms during which participants could blink. E-Prime Studio version 2 (Psychology Software Tools Inc., PA, USA) was used to present stimuli and record participants' responses.

Auditory stimuli were delivered binaurally through MEG-compatible ER3A insert earphones (Etymotic Research Inc., IL, USA). There was a 26ms ± 2ms delay in sound delivery due to the transmission of auditory signal from the stimulus computer to the earphones. This sound delivery delay was corrected in the following analyses. To ensure that participants were able to hear the stimuli through both earphones, a short hearing test was conducted before the main experiment.

Sentence continuation pre-tests

To obtain human incremental interpretations for each of the HiTrans and LoTrans sentences (120 in total), we conducted two continuation pre-tests which involved two different groups of native British English speakers (30 participants in the first pre-test, 18 participants in the second, aged between 18 and 34 years) who did not participate in the main experiment.

Specifically, participants wore headphones and were seated in front of a computer. They listened to a fragment of one of the HiTrans/LoTrans sentences starting from its onset and continuing until a certain position in the sentence, and then they were asked to complete this sentence by producing a meaningful continuation. The sentence fragment was binaurally presented up to the Verb1 (e.g., *The dog found...*) in the first pre-test, and was presented up to the end of the prepositional phrase in the second pre-test (e.g., *The dog found in the park...*). In the second pre-test, participants were allowed to provide a full stop as a continuation if they thought that what they had heard was a complete sentence.

Based on the continuations obtained in the first pre-test, we calculated the probability of direct object or prepositional phrase in the continuations immediately after the Verb1, i.e., DO probability and PP probability. This provided contextualized measures of the transitive or the intransitive use of Verb1 given the preceding subject noun phrase. Specifically, we defined Verb1 transitivity as DO probability/(1- DO probability) and defined Verb1 intransitivity as (1-DO probability)/DO probability. Given the continuations collected in the second pre-test, we calculated the probability of a main verb in the continuations immediately after the prepositional phrase, i.e., MV probability, which directly reflected a listener's structural interpretation by the end of the prepositional phrase. The absence of a main verb in the continuation after the prepositional phrase indicated that the Active interpretation was taken and the Verb1 was considered as the "main verb", whereas the appearance of a main verb in the continuation indicated that the Passive interpretation was taken (i.e., Verb1 was interpreted as a passive verb), and thus a main verb was needed to complete the sentence.

EMEG and MRI acquisition

Participants were seated in a magnetically shielded room (IMEDCO GMBH, Switzerland) with their head placed in the helmet of the MEG scanner. MEG data were collected using a Neuromag Vector View system (Elekta, Helsinki, Finland) with 102 magnetometers and 204 planar gradiometers at 1kHz sampling rate. Simultaneous EEG was recorded at 1kHz sampling rate from 70 Ag-AgCl electrodes within an elastic cap (ESACYCAP GmbH, Herrsching-Breitbrunn, Germany). Vertical and horizontal eye movements were recorded by two EOG electrodes attached below and lateral to the left eye, cardiac signals were recorded by two ECG electrodes attached separately to the right shoulder blade and left torso. Five head position indicator (HPI) coils were used to monitor head motion. A 3D digitizer was used to record the position of EEG electrodes, HPI coils and head points on participants' scalp relative to the 3 anatomical fiducials (i.e., nasion and bilateral preauricular points). To source localize EMEG data, T1-weighted MPRAGE structural MRI image with 1mm isotropic resolution was acquired using a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). All EMEG and MRI data were collected at the MRC Cognition and Brain Sciences Unit, University of Cambridge.

EMEG preprocessing and source localization

Maxfilter (Elekta, Helsinki, Finland) was applied to raw MEG data for bad channel removal and head-motion compensation. Signals outside the brain were removed using the temporal extension of signal-space separation (Taulu and Simola 2006). EMEG data were low-pass

filtered at 40 Hz and high-pass filtered at 0.5 Hz with a 5th order bidirectional Butterworth filter using SPM12 (Wellcome Trust Centre for Neuroimaging, UCL). Independent component analysis (ICA) was conducted using EEGLAB (SCCN, UCSD), components related to blink, eye-movement and physiological noises were removed according to the correlation with EOG, ECG signals and further visual inspection. The preprocessed EMEG data were then downsampled to 200 Hz. Three epochs were extracted from the continuous EMEG recordings of each HiTrans or LoTrans sentence with auditory delivery delay corrected - V1 epoch was aligned to the onset of the Verb1, PP1 epoch was aligned to the onset of the preposition, MV epoch was aligned to the onset of the main verb. All the three epochs were 600ms in length. For all three epochs, baseline correction was performed using the signal from a silent period (i.e., -200 ms to 0 ms relative to sentence onset). Finally, automatic artefact rejection was conducted to exclude trials with signals that exceeded pre-defined amplitude thresholds (60 ft/mm for gradiometers, 3000 ft for magnetometers and 200 uV for EEG electrodes). The uniqueness point of V1/PP1/MV was defined as the earliest point in time when this word can be fully recognized after removing all of its phonological competitors. We first identified the phoneme by which this word can be uniquely recognized according to CELEX (Baayen *et al.* 1993). Then, we manually labelled the offset of this phoneme in the auditory file of the spoken sentence.

EMEG data source localization was performed using SPM12. Source space was modelled by a cortical mesh consisting of 8196 vertices. The sensor positions were co-registered to individual T1-weighted structural image by aligning fiducials and the digitized head shape to the outer scalp mesh. MEG forward model was constructed using the single-shell model (Sarvas 1987), while EEG forward model was built using the boundary element model (Mosher et al. 1999). Inversion of EMEG data was conducted for V1, PP1 and MV epochs separately using the least-squares minimum norm method (Hamalainen and Ilmoniemi 1994) and an empirical Bayesian MEG and EEG data fusion scheme implemented in SPM12 (Henson et al. 2009).

Incremental structural representations of BERT

To obtain incremental structural representations of BERT, we adopted a structural probing approach (Hewitt and Manning 2019) to quantify a sentence's structure by estimating each word's parse depth in the corresponding dependency parse tree based on the contextualized word embeddings from the hidden states of BERT, which explicitly considers the specific contents of the words in the input. Specifically, a structural probing model was trained to find an optimal linear transformation to be applied to the BERT contextualized embeddings of words in the input sentence, so that the squared L2 norm of the transformed word embeddings provided the best estimate for each word's parse depth of in the dependency parse tree of this sentence.

We followed the procedure described in a previous study (Hewitt and Manning 2019) and trained a structural probing model for each BERT layer with the annotated corpus from Penn Treebank (Marcus et al. 1993). Contextualized word embeddings were extracted from each of the 24 layers in a pre-trained version of BERT (BERT-large-cased) using HuggingFace (Wolf

et al. 2019). For each BERT layer, the training process was repeated 10 times with different random initializations, the averaged BERT parse depth was used in the following analyses. The performance of structural probing models trained by different BERT layers was evaluated by root accuracy. Root accuracy is defined as the percentage of the sentences in which the samllest parse depth is assigned to the main verb (i.e., the root of the dependency parse tree has a parse depth of 0) when the whole sentence is input to the model.

Each HiTrans or LoTrans sentence was input word-by-word to the trained BERT structural probing models, which resulted in a vector consisting of the parse depth of each word in the incremental input (e.g., a 3-dimensional BERT parse depth vector for the input "*The dog found*..."). Taking advantage of the bi-directional attention mechanism of BERT, the parse depth of each preceding word was constantly updated as the input unfolded word-by-word, capturing the incrementality of speech comprehension. Besides, we defined interpretive mismatch as the cosine distance between an incremental BERT parse depth vector and the corresponding incremental context-free parse depth vector for the Passive or the Active interpretation. The smaller the interpretive mismatch with one particular interpretation, the higher the preference for this interpretation given BERT structural representations.

To determine the contribution of the words at different positions in a sentence to the incremental BERT parse depth vectors, we shuffled the parse depths of the words at a particular position across sentences at a time and kept the parse depths of the other words unchanged. Then we calculated the Spearman distance (i.e., 1-Spearman's rho) between the original BERT parse depth vector and the shuffled counterpart. The higher this distance, the more important the words at this position are to the BERT parse depth vectors (see **Appendix 1-figure 3**).

Corpus-based measures of multifaceted constraints and their interpretative coherence

We quantified subject noun thematic role properties and Verb1 transitivity preference based on a concatenated corpus (3.4 billion tokens, 162.1 million sentences) consisting of the British National Corpus, the Wikipedia dump (by October 2020) and ukWaC (Baroni et al. 2009). A dependency parser (Mrini et al. 2020) was first applied to each sentence in the concatenated corpus to specify the subject noun and the verb(s) related to it. For the subject noun in each sentence, we used a semantic role labelling model (Li et al. 2020) to obtain its thematic role. For simplicity, we only considered the thematic role of an agent or a patient (Dowty 1991).

For each subject noun in HiTrans and LoTrans sentences, we counted separately how many times it took the thematic role of an agent or a patient in the concatenated corpus. Then we defined its agenthood as the ratio of the number of its appearances as an agent to that of its appearances as a patient, and versa visa for its patienthood. We also counted separately how many times the Verb1 in each HiTrans or LoTrans sentence took a direct object or alternative SCFs in the corpus. Each Verb1's transitivity was defined as the ratio of the frequency it took a direct object to that it took alternative SCFs, and vice versa for its intransitivity. By doing so, we obtained subject noun agenthood and patienthood, Verb1 transitivity and intransitivity. Note that the Verb1 (in)transitivity estimated from the human continuations in the pre-test is context-dependent given the specific preceding subject noun, while the Verb1 (in)transitivity

estimated from the concatenated corpus is context-independent in the sense that it accounted for every appearance of this verb without being biased to a specific context.

We further derived Passive/Active index capturing the interpretative coherence between the subject noun and Verb1 as they affected Passive and Active interpretations separately. The Passive index was obtained by multiplying subject noun patienthood with Verb1 transitivity, given that both high subject noun patienthood and high Verb1 transitivity coherently prefer a Passive interpretation as the prepositional phrase is heard (e.g., *The **dog found** in the park...*). In contrast, the Active index was obtained by multiplying subject noun agenthood by Verb1 intransitivity, capturing the preference for an Active interpretation (e.g., *The **king walked** in the garden...*). Besides, we also calculated a contextualized version Passive/Active index by using Verb1 (in)transitivity derived from human continuation pre-tests instead of that derived from the concatenated corpus. In addition, we derived a non-directional index by applying logarithmic transformation to the ratio measures of lexical constraints (i.e., subject noun agenthood or patienthood, Verb1 intransitivity or transitivity) before multiplying them. This manipulation removed the directionality of the Passive or the Active index, so that the non-directional index only indicates the interpretative coherence between the subject noun and Verb1 regardless of which interpretation is considered (see illustrations of directionality in **Appendix 1-figure 17**).

Spatiotemporal searchlight representational similarity analysis (ssRSA)

ssRSA was conducted to compare the (dis)similarity structure of BERT structural measures or the multifaceted probablistic constraints with the (dis)similarity structure of observed spatiotemporal patterns of listeners' brain activity. We used a spatiotemporal searchlight with a 10mm spatial radius and 30ms temporal radius (i.e., a 60 ms sliding time window) which was mapped across the whole brain in the source-localized EMEG.

For brain activity, we constructed data representational dissimilarity matrix (RDM) by vectorizing the source-localized EMEG data within each spatiotemporal searchlight for all the trials (i.e., 60 HiTrans sentences and 60 LoTrans sentences) and calculated the pairwise Pearson's correlation distance (i.e., 1 - Pearson's r) among them, which resulted in a 120 x 120 data RDM. Multivariate normalisation was applied to improve the reliability of distance measures and reduce the task-irrelevant heteroscedastic structure across trials and vertices (Guggenmos et al. 2018). Model RDMs of the same size (i.e., 120 x 120) were constructed by calculating either the absolute pair-wise difference for a scalar measure (e.g., SN agenthood/patienthood, Passive/Active index, BERT Verb1 parse depth) or the cosine distance among the incremental BERT parse depth vectors of the 120 sentences (see **Appendix 1-table 1** for a summary of all the model RDMs). We used ratio measures to represent subject noun agenthood/patienthood, Verb1 transitivity/intransitivity and Passive/Active index because they provided the directionality needed to differentiate the two opposite aspects of the same lexical constraint in the model RDMs (see illustrations in **Appendix 1-figure 11**) and made it possible to test them separately in the brain.

Each model RDM was compared against the data RDM of a searchlight centered at each vertex and time point using Spearman's rank correlation, which resulted in a time-series of model fit (i.e., rank correlation coefficient rho) for each vertex. For each time point, a one-tailed one-sample t-test was conducted at each vertex with the fits of all participants for this model RDM to test whether the mean model fit is significantly above zero. Cluster permutation tests were performed for multiple comparison correction with 5,000 nonparametric permutations, vertex-wise $P < 0.01$ and cluster-wise $P < 0.05$.

Granger causality analysis (GCA) based on ssRSA model fits

GCA was conducted to investigate the relationship between multifaceted constraints and BERT structural measures in terms of their effects in the brain, i.e., ssRSA model fits of the corresponding model RDMs. For a given model RDM, each participant's non-thresholded whole-brain model fit time-series were normalized and concatenated across participants, which resulted in a vertex by time-point matrix. Non-negative matrix factorization (NMF) was applied to this concatenated model fit matrix with negative model fits zeroed. NMF was repeated 20 times with random starting values using a multiplicative update algorithm in MATLAB, results with the least root mean square residual (RMS) was used in the following analyses. The optimal number of NMF factors was determined by searching for the one with the least RMS in a wide range of factor numbers (from 2 to 50). With the optimal number of factors, the resulted time-series of NMF factors from all participants were reshaped into a factor by time-point by participant matrix which was used as the input of multivariate GCA implemented by the Multivariate GCA toolbox (Barnett and Seth 2014). Multivariate GCA was conducted using two sets of NMF factors derived from the fits of two model RDMs with Akaike information criterion (AIC) adopted for GCA model order estimation. GC significance was determined by a permutation test in which 1,000 surrogate data sets were created by randomly rearranging short time windows (of length model order) from the original factor time-course. *P*-values below 0.005 were refined via a tail approximation from the Generalised Pareto Distribution using PALM (Winkler et al. 2016). Multiple comparisons were controlled with false discovery rate (FDR) alpha < 0.05 for both between- and within-model RDM GC connections.

Acknowledgements

This research was funded by European Research Council Advanced Investigator Grant to L.K.T. under the European Community's Horizon 2020 Research and Innovation Programme (2014-2022 ERC Grant Agreement 669820). B.L. was supported by the Ministry of Science and Technology of China and Changping Laboratory. We thank Billi Randall and Barry Devereux for their valuable contributions to early experimental design and to stimulus development; and Hun S. Choi, Benedict Vassileiou, John Hewitt, Tao Li, Yi Zhu, Nai Ding and Giorgio Marinato for helpful discussions.

Author contributions

Conceptualization: L.K.T., W.D.M., B.L.
Investigation, Data curation: B.L., Y.F.
Methodology, Formal Analysis & Visualization: B.L.
Funding acquisition & Project administration: L.K.T.
Supervision: L.K.T., W.D.M.
Writing – original draft, review & editing: B.L., W.D.M., L.K.T.

Declaration of interests

Authors declare no competing interests.

Data and materials availability

Preprocessed E/MEG data and scripts are available in the following repository https://osf.io/7u8jp/.

References

Altmann GT. 1998. Ambiguity in sentence processing. Trends in Cognitive Sciences. 2:146-152.

Altmann GTM, Ekves Z. 2019. Events as intersecting object histories: A new theory of event representation. Psychological Review. 126:817-840.

Baayen RH, Piepenbrock R, van H R. 1993. The {CELEX} lexical data base on {CD-ROM}.
Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. 2017. Discovering Event Structure in Continuous Narrative Perception and Memory. Neuron. 95:709-721 e705.

Bao P, She L, McGill M, Tsao DY. 2020. A map of object space in primate inferotemporal cortex. Nature. 583:103-108.

Barnett L, Seth AK. 2014. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. Journal of Neuroscience Methods. 223:50-68.

Baroni M, Bernardini S, Ferraresi A, Zanchetta E. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation. 43:209-226.

Bengio Y, Lecun Y, Hinton G. 2021. Deep Learning for AI. Communications of the ACM. 64:58-65.

Bever TG. 1970. The cognitive basis for linguistic structures. In: Hayes JR, editor. Cognition and the Development of Language New York: John Wiley.

Bicknell K, Elman JL, Hare M, McRae K, Kutas M. 2010. Effects of event knowledge in processing verbal arguments. Journal of Memory and Language. 63:489-505.

Binz M, Schulz E. 2023. Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences of the United States of America. 120:e2218523120.

Bisk Y, Holtzman A, Thomason J, Andreas J, Bengio Y, Chai J, Lapata M, Lazaridou A, May J, Nisnevich A, Pinto N, Turian J editors. Experience Grounds Language; 2020 November; Online:Association for Computational Linguistics. 8718-8735 p.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. 2020. Language models are few-shot learners. Advances in neural information processing systems. 33:1877-1901.

Caucheteux C, Gramfort A, King JR. 2022. Deep language algorithms predict semantic comprehension from brain activity. Scientific Reports. 12:16327.

Caucheteux C, Gramfort A, King JR. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour. 7:430-441.

Caucheteux C, King JR. 2022. Brains and algorithms partially converge in natural language processing. Communications Biology. 5:134.

Choi HS, Marslen-Wilson WD, Lyu B, Randall B, Tyler LK. 2021. Decoding the Real-Time Neurobiological Properties of Incremental Semantic Interpretation. Cereb Cortex. 31:233-247.

de Marneffe M-C, MacCartney B, Manning CD editors. Generating typed dependency parses from phrase structure parses, Proceedings of the 5th International Conference on Language Resources and Evaluation; 2006 May 22-28, 2006; Genoa, Italy:European Language Resources Association. 449-454 p.

Devlin J, Chang M-W, Lee K, Toutanova K editors. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 June 2-7, 2019; Minneapolis, MN, USA:Association for Computational Linguistics. 4171-4186 p.

Doerig A, Sommers RP, Seeliger K, Richards B, Ismael J, Lindsay GW, Kording KP, Konkle T, van Gerven MAJ, Kriegeskorte N, Kietzmann TC. 2023. The neuroconnectionist research programme. Nature Reviews Neuroscience. 24:431-450.

Donhauser PW, Baillet S. 2019. Two Distinct Neural Timescales for Predictive Speech Processing. Neuron.

Dowty D. 1991. Thematic Proto-Roles and Argument Selection. Language. 67:547-619.

Duncan J. 2010. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends in Cognitive Sciences. 14:172-179.

Elman JL. 1990. Finding Structure in Time. Cognitive Science. 14:179-211.

Elman JL. 1993. Learning and development in neural networks: the importance of starting small. Cognition. 48:71-99.

Everaert MBH, Huybregts MAC, Chomsky N, Berwick RC, Bolhuis JJ. 2015. Structures, Not Strings: Linguistics as Part of the Cognitive Sciences. Trends in Cognitive Sciences. 19:729-743.

Frazier L. 1987. Syntactic processing: evidence from Dutch. Natural Language & Linguistic Theory. 5:519-559.

Frazier L, Rayner K. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology. 14:178-210.

Friederici AD. 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends in Cognitive Sciences. 16:262-268.

Friederici AD, Bahlmann J, Heim S, Schubotz RI, Anwander A. 2006. The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proceedings of the National Academy of Sciences of the United States of America. 103:2458-2463.

Giordano BL, Esposito M, Valente G, Formisano E. 2023. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. Nature Neuroscience.

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, Jansen A, Gazula H, Choe G, Rao A, Kim C, Casto C, Fanda L, Doyle W, Friedman D, Dugan P, Melloni L, Reichart R, Devore S, Flinker A, Hasenfratz L, Levy O, Hassidim A, Brenner M, Matias Y, Norman KA, Devinsky O, Hasson U. 2022. Shared computational principles for language processing in humans and deep language models. Nature Neuroscience. 25:369-380.

Guggenmos M, Sterzer P, Cichy RM. 2018. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. Neuroimage. 173:434-447.

Hamalainen MS, Ilmoniemi RJ. 1994. Interpreting magnetic fields of the brain: minimum norm estimates. Medical & Biological Engineering & Computing. 32:35-42.

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, de Lange FP. 2022. A hierarchy of linguistic predictions during natural language comprehension. Proceedings of the National Academy of Sciences of the United States of America. 119:e2201968119.

Henson RN, Mouchlianitis E, Friston KJ. 2009. MEG and EEG data fusion: simultaneous localisation of face-evoked responses. Neuroimage. 47:581-589.

Hewitt J, Liang P editors. Designing and interpreting probes with control tasks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 November 3-7, 2019; Hong Kong, China:Association for Computational Linguistics. 2733-2743 p.

Hewitt J, Manning CD editors. A structural probe for finding syntax in word representations, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 June 2-7, 2019; Minneapolis, MN, USA:Association for Computational Linguistics. 4129-4138 p.

Humphreys GF, Lambon Ralph MA, Simons JS. 2021. A Unifying Account of Angular Gyrus Contributions to Episodic and Semantic Cognition. Trends in Neurosciences. 44:452-463.

Jackendoff R, Jackendoff RS. 2002. Foundations of language: Brain, meaning, grammar, evolution: Oxford University Press, USA.

Jung-Beeman M. 2005. Bilateral brain processes for comprehending natural language. Trends in Cognitive Sciences. 9:512-518.

Jurayj W, Rudman W, Eickhof C editors. Garden Path Traversal in GPT-2, Proceedings of the 5th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP; 2022 December 8, 2022; Abu Dhabi, United Arab Emirates:Association for Computational Linguistics. 305-313 p.

Kietzmann TC, Spoerer CJ, Sorensen LKA, Cichy RM, Hauk O, Kriegeskorte N. 2019. Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences of the United States of America. 116:21854-21863.

Korhonen A, Krymolowski Y, Briscoe T editors. A large subcategorization lexicon for natural language processing applications, Proceedings of International Conference on Language Resources and Evaluation; 2006; Genoa, Italy. 1015-1020 p.

Kriegeskorte N, Douglas PK. 2018. Cognitive computational neuroscience. Nature Neuroscience. 21:1148-1160.

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience. 2:4.

Kuperberg GR. 2007. Neural mechanisms of language comprehension: challenges to syntax. Brain Research. 1146:23-49.

Law R, Pylkkanen L. 2021. Lists with and without Syntax: A New Approach to Measuring the Neural Processing of Syntax. Journal of Neuroscience. 41:2186-2196.

Li T, Jawale PA, Palmer M, Srikumar V editors. Structured Tuning for Semantic Role Labeling, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online.

Linzen T, Baroni M. 2021. Syntactic Structure from Deep Learning. Annual Review of Linguistics. 7:195-212.

Lyu B, Choi HS, Marslen-Wilson WD, Clarke A, Randall B, Tyler LK. 2019. Neural dynamics of semantic composition. Proceedings of the National Academy of Sciences of the United States of America. 116:21318-21327.

MacDonald MC, Pearlmutter NJ, Seidenberg MS. 1994. The lexical nature of syntactic ambiguity resolution. Psychological Review. 101:676-703.

Mahowald K, Diachek E, Gibson E, Fedorenko E, Futrell R. 2023. Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. Cognition. 241:105543.

Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences of the United States of America. 117:30046-30054.

Marcus MP, Santorini B, Marcinkiewicz MA. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics. 19:313-330.

Marslen-Wilson W, Tyler LK. 1980. The temporal structure of spoken language understanding. Cognition. 8:1-71.

Marslen-Wilson WD. 1975. Sentence perception as an interactive parallel process. Science. 189:226-228.

Marslen-Wilson WD, Tyler LK, Koster C. 1993. Integrative Processes in Utterance Resolution. Journal of Memory and Language. 32:647-666.

Matchin W, Hickok G. 2020. The cortical organization of syntax. Cereb Cortex. 30:1481-1498.
McClelland JL, Hill F, Rudolph M, Baldridge J, Schutze H. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. Proceedings of the National Academy of Sciences of the United States of America. 117:25966-25974.

McRae K, Matsuki K. 2013. Constraint-based models of sentence processing. Sentence processing. 519:51-77.

Metusalem R, Kutas M, Urbach TP, Elman JL. 2016. Hemispheric asymmetry in event knowledge activation during incremental language comprehension: A visual half-field ERP study. Neuropsychologia. 84:252-271.

Metusalem R, Kutas M, Urbach TP, Hare M, McRae K, Elman JL. 2012. Generalized event knowledge activation during online sentence comprehension. Journal of Memory and Language. 66:545-567.

Mosher JC, Leahy RM, Lewis PS. 1999. EEG and MEG: forward solutions for inverse methods. IEEE Transactions on Biomedical Engineering. 46:245-259.

Mrini K, Dernoncourt F, Tran QH, Bui T, Chang W, Nakashole N editors. Rethinking Self-Attention: Towards Interpretability in Neural Parsing; 2020:Association for Computational Linguistics. 731-742 p.

Nelson MJ, El Karoui I, Giber K, Yang X, Cohen LD, Koopman HJ, Cash SS, Naccache L, Hale JT, Pallier C, Dehaene S. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. Proceedings of the National Academy of Sciences of the United States of America. 114:E3669 - E3678.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems.

Pallier C, Devauchelle A-D, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. Proceedings of the National Academy of Sciences of the United States of America. 108:2522-2527.

Pavlick E. 2022. Semantic Structure in Deep Learning. Annual Review of Linguistics. 8:447-471.

Rabovsky M, Hansen SS, McClelland JL. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nature Human Behaviour. 2:693-705.

Sarvas J. 1987. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Physics in Medicine & Biology. 32:11-22.

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences of the United States of America. 118:e2105646118.

Sheahan H, Luyckx F, Nelli S, Teupe C, Summerfield C. 2021. Neural state space alignment for magnitude generalization in humans and recurrent networks. Neuron. 109:1214-1226 e1218.

Smallwood J, Bernhardt BC, Leech R, Bzdok D, Jefferies E, Margulies DS. 2021. The default mode network in cognition: a topographical perspective. Nature Reviews Neuroscience. 22:503-513.

Taulu S, Simola J. 2006. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Physics in Medicine & Biology. 51:1759-1768.

Tenney I, Xia P, Chen B, Wang A, Poliak A, McCoy RT, Kim N, Van Durme B, Bowman SR, Das D editors. What do you learn from context? probing for sentence structure in contextualized word representations, the 7th International Conference on Learning Representations; 2019 May 6-9, 2019; New Orleans, LA, USA.

Toneva M, Mitchell TM, Wehbe L. 2022. Combining computational controls with natural text reveals aspects of meaning composition. Nature Computational Science. 2:745-757.

Troyer M, McRae K, Kutas M. 2022. Wrong or right? Brain potentials reveal hemispheric asymmetries to semantic relations during word-by-word sentence reading as a function of (fictional) knowledge. Neuropsychologia. 170:108215.

Trueswell JC, Tanenhaus MK. 1994. Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. In. Perspectives on sentence processing Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc p 155-179.

Tyler LK, Marslen-Wilson WD. 1977. The On-Line Effects of Semantic Context on Syntactic Processing. Journal of Verbal Learning and Verbal Behavior. 16:683-692.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. Advances in neural information processing systems. 30.

Winkler AM, Ridgway GR, Douaud G, Nichols TE, Smith SM. 2016. Faster permutation inference in brain imaging. Neuroimage. 141:502-516.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf Re, Funtowicz M, Brew J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv. abs/1910.03771.

Yamins DL, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience. 19:356-365.

Yang GR, Joglekar MR, Song HF, Newsome WT, Wang XJ. 2019. Task representations in neural networks trained to perform many cognitive tasks. Nature Neuroscience. 22:297-306.

Yeshurun Y, Nguyen M, Hasson U. 2021. The default mode network: where the idiosyncratic self meets the shared social world. Nature Reviews Neuroscience. 22:181-192.